uc3m | Universidad Carlos III de Madrid

University Degree in Industrial Technologies Engineering
Academic Year: 2017-1018

Bachelor Thesis

"Smart-Real-Estate: Design and implementation of a system for calculating house prices"

Author:
Ana de Lázaro Noreña

Tutor:
Dr. José María Álvarez Rodríguez

Leganés, June 19th

SUMMARY

The Big Data world has, in the last few years, exploded, the reason for this is that the amount of information nowadays available is exponentially growing every day. The purpose of this project is to use this available data to fix a universally shared problem when buying a house. To be able to know if the house is over or under priced useful for both the seller and the buyer as it will give them an idea on how to proceed.

To solve this, the present project is going to build a logical model to be able to either classify the expected price in different categories or to predict the exact value. The only needed values to do so would be the attributes of the home: characteristics both inside the home and outside.

Along this project, research on the different logical models and algorithms can be found, as well as the engineering process to solve the problem: How much should this house be worth?

This engineering process will consist of three stages: Analysis, Design and Implementation.

Finally, a short study on the future enhancements is done as this project is considered as the start of what could be a model to classify every house in Madrid but for now it will be for a District.

Key words: Logical models, algorithm, real estate, classification, Big Data, attributes.

ACKNOWLEDGMENTS

First and most important I would like to thank my advisor, Dr. José María Álvarez Rodríguez who patiently helped me throughout this process. I know it wasn't easy as my base knowledge on programming and logical models was close to none but still he pushed and motivated me to carry on.

This project marks an end of my degree and therefore, I would like to take the opportunity to thank my family and university classmates and friends. My family has been there to motivate and listen to me while my classmates have made the most difficult subjects achievable by helping every chance they got.

Thank you all, it can't be said enough.

TABLE OF CONTENTS

TABLE OF FIGURES

IMAGES

GRAPHS

SCHEMES

TABLES

CHAPTER 1: MOTIVATION

Being part of the Business Intelligence (BI) department at a start-up company for a year introduced me to the world of Big Data Analytics. Never before had I realize the amount of data accessible to anyone and the amount of predictions and models that could be created with the right use of it.

Building a model such as this one meant two things: the knowledge acquired on logical models and the better understanding of the housing market. Although I hadn't studied informatics since my first year at Industrial Engineering, programing was what most called to me and the best way of learning is by starting from scratch, being able to build something using not only different tools such as Anaconda but also programming languages such as R and Python.

By taking a look at different big data competitions the idea of trying to do a model for estimating the price of homes was born (Kaggle, 2017). This competition gave a 1,2 million dollars to whoever was able to improve the Zillow zestimate which gave me the idea of trying to create a new model for estimating house prices.

On the other hand, the housing market is not only a huge market but although there is a lot of data to be collected there aren't many models to try to predict the real market prices, the reasons for this will be explained in Chapter 3: State of the art.

From the start of the project there was the distinct possibility that the results wouldn't be what were expected and, therefore, the model wouldn't be useful. The data is not exact and the house pricing market has many variables that won't be easily found or predicted. Also, it is a segmented market and sometimes two buildings have high differences in price even if they are next to each other, making the predictions much more difficult.

There are models being created all the time but most of them are not able to differentiate between same building houses, giving the same value to every floor and every apartment. The idea of trying to create something different and hopefully better than the BBVA model to predict house prices was a great challenge that motivated me throughout this project.

CHAPTER 2: OBJECTIVES

- To understand the cadastral value and its implications.
- To analyse and research the factors that affect the price of any house/apartment in a district of Madrid, 'Barrio de Salamanca'.
- To find the value of those factors in different databases.
- To automatize as much as possible the access to those databases.
- To be able to see which attributes do affect the value and which don't.
- To create a Logical Model by trying different methods, comparing and analysing them in order to decide which one is the better option.
- To analyse and discuss the results.
- To analyse the future improvements of the model and the implications of using it both in the future and for an entire city.

CHAPTER 3: STATE OF THE ART

This chapter will take a look at the state of the art of logical models, the housing market in Spain, the cadastral value and the different housing portals for buying and selling houses.

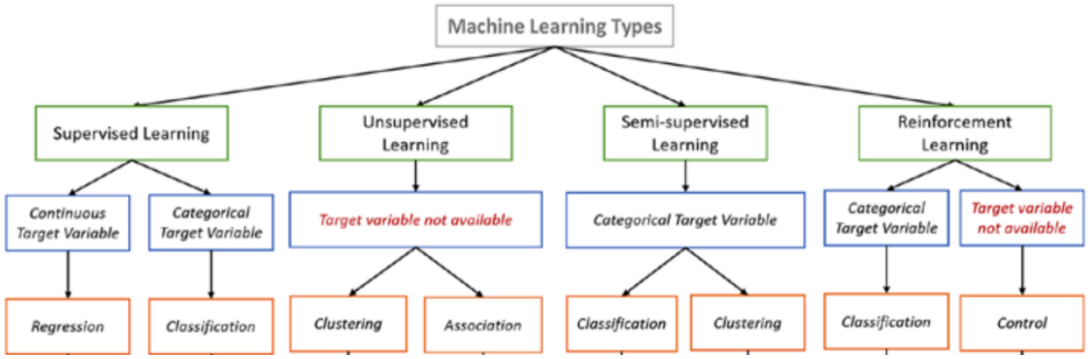
3.1 Logical models

A logical model is a way to describe and understand the relationship between different elements/variables. They can be both used to describe what is planned and what is expected to happen.

Although they are used by both the private and the public sector they have recently received an increased interest by all sectors as a way to make decisions as they enhance learning through all the data available. They are used from describing a district's education improvement plan to predicting housing prices (Fumo, 2017).

During this project, the modelling will be done by using WEKA which provides the user with more than 20 different types of algorithms. The different types of algorithms can be classified in different ways. This is one of them:

SCHEME I: DIFFERENT MACHINE LEARNING TYPES (Fumo, 2017)



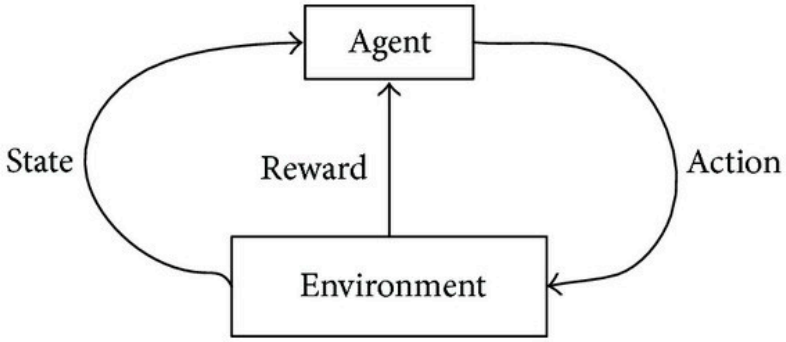
Supervised learning basically is that the algorithm is trained and finds the `function` that best describes the input of data to get to the output. They model relationships and dependencies between the target prediction and the input variables. The objective is to use those relationships to predict what is needed from the input used (Fumo, 2017).

Unsupervised learning is used more for pattern detection and descriptive modelling as it uses unlabelled data. These algorithms have as an objective to both describe the data better to users and help in finding meaningful insights into the data (Fumo, 2017).

Semi-supervised Learning is a mix of the previous two, they use both labelled and unlabelled data. It is the preferred method when huge amounts of data are involved as labelling needs human experts which will increase the cost (Fumo, 2017).

Reinforcement learning has as its objective to use observations gathered from the interaction with its surroundings to either maximize or minimize the reward and the risk. They allow machines and software agents to maximize its performance (Simonini, 2018).

IMAGE I: UNSUPERVISED LEARNING (Fumo, 2017)



Clearly, in this project the model will fall into the category of supervised learning as all the data is labelled. Both regression and classification will be tried and compared to each other although the research shows regression as the commonly one used.

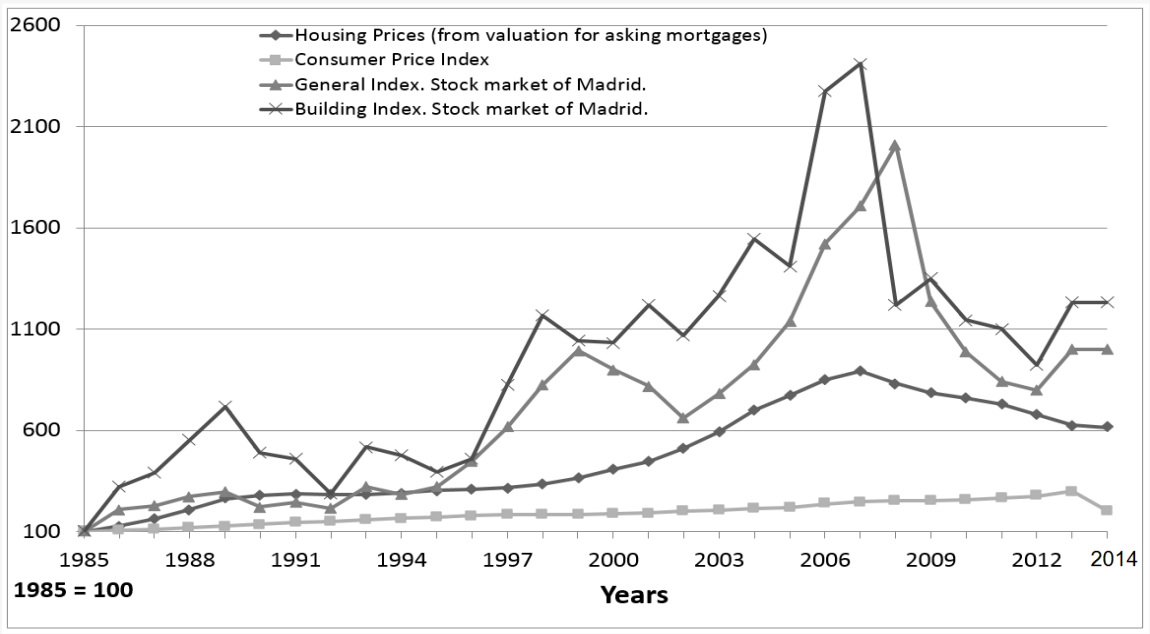
3.2 Housing market

The Spanish housing market has a distinct intensity which means an instability of prices within the city districts themselves, there is a need to find an approximation of the real market value of the houses (Rodríguez López, 2006).

Every time there is an economical transaction of a house this produces new prices of the rest, affecting more the ones closer but in some level upsetting the whole housing market. Furthermore, the dynamics of supply and demand that make the prices seem random have their bases in the urgency, necessity and desire of the buyer.

By researching the housing prices since 1985 till 2014 the following graph of housing prices (€/m^2) was created:

GRAPH I: HOUSE PRICING PER SQUARE METER FROM 1985-2014 (Tasación)



While the bigger ups and downs can be more or less explained by looking at what could have affected them the smaller ones can't.

Some explanation by researchers and professionals in the field show that some of the ups coincide with the following scenarios: Credit market liberalization, the rise in population and the introduction of the euro. While some of the downs coincide with: the household credit was cut and the economic crisis. Those are cut and dry factors that everyone knows will affect the price but there is a need to take a look at what affect the prices when we compare them at a moment in time to each other (Rodríguez López, 2006).

The variables that will be taken into account and the difficulties getting them are explained in Chapter four.

3.3 Cadastral value

This characteristic already takes into account all the other factors but because of its importance in the real value of a house it deserves a further look by itself. There are multiple thesis and research projects that take a deep look at this value and where it comes from but for the purpose of this project there is no need to go that far, it will just be a summary of how it is calculated and where it comes from.

The system used at the moment to determine the cadastral value of urban housing in a district it is essentially based on a collective valorisation that has its origins in the

approval of presented cadastral values the last available ones were used for our district Barrio de Salamanca, the ones from 2011 (Appendix I).

Once the values are approved a formula is used to calculate the price of each flat. In the land registry databases, all the characteristics of each flat are recorded and any change such as dividing a house in two or even merging two rooms into one must be registered there (Ministerio de Hacienda).

The formula of the cadastral value is the sum of the cadastral value of the area built inside the exterior line of the perimeter walls and the value of the construction, all multiplied by the localization factor. The terraces compute as a 50% as long as they are not covered by three out of four walls, otherwise they will compute as 100% (DCC, 2017).

The cadastral value of the land (Vs):

$$Vs = Area_{built} x\ Vr\ x\ Correction\ factors$$

Vr is the repercussion value and can be found in Appendix I.

The cadastral value of the construction (Vc):

$$Vc = Area_{built} x\ typology\ value\ x\ correction\ factors$$

The typology value for our project is the home value of each value zone, both found in Appendix I.

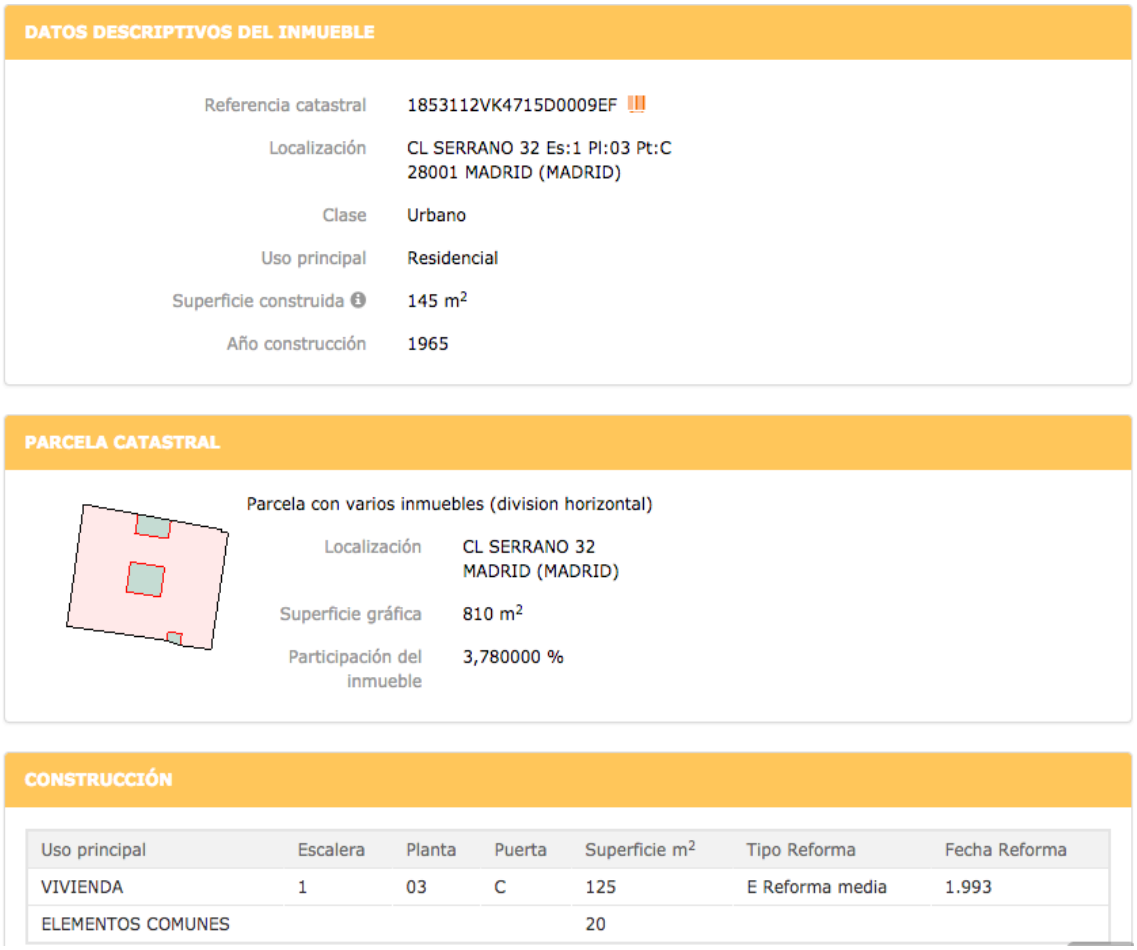
Then the Cadastral value is the sum of Vs and Vc multiplied by the localization value, which can also be found in Appendix I.

The correction factors are seven and they take into account from the number of walls that show to the street, the number of floors of the building, the length of the building to the protection of the land. While for Vs they look at the land of the whole building, Vc takes a deeper look to the exact characteristics of the apartment (DCC, 2017).

Although the cadastral values are supposed to be of free access to the residents only the cadastral value of your own home can be accessed (Certicalia). For the purpose of this project only an approximation of them was made.

This is an example of the information allowed to the public:

IMAGE II: EXAMPLE OF PUBLIC CADASTRE INFORMATION (Gobierno de España)



DATOS DESCRIPTIVOS DEL INMUEBLE

Referencia catastral	1853112VK4715D0009EF
Localización	CL SERRANO 32 Es:1 Pl:03 Pt:C 28001 MADRID (MADRID)
Clase	Urbano
Uso principal	Residencial
Superficie construida	145 m²
Año construcción	1965

PARCELA CATASTRAL

Parcela con varios inmuebles (division horizontal)

Localización	CL SERRANO 32 MADRID (MADRID)
Superficie gráfica	810 m²
Participación del inmueble	3,780000 %

CONSTRUCCIÓN

Uso principal	Escalera	Planta	Puerta	Superficie m²	Tipo Reforma	Fecha Reforma
VIVIENDA	1	03	C	125	E Reforma media	1.993
ELEMENTOS COMUNES				20		

3.4 Housing portals

For the purpose of this project there was a need to get as much data as available from a housing portal in order to find all the variables needed.

Since the era of technology, apps and internet exploded, real estate agents are recommended and used less as their cut in the sell and buy transactions of a house is much higher than in the housing portals. Also, the housing portals are able to recollect and use much more data than the real estate agencies. On the other hand, for the interested buyer there are more offers on housing portals and they are easier to sort through. Even real estate agencies use housing portals such as Idealista to advertise their clients' flats.

All around the world, every day there are new housing portals created, but each has its strengths. For example, Air BnB is used for short term renting of rooms and flats, while the Idealista is better at long term renting and buying and selling homes and flats.

So, after researching the most popular housing portals in Spain, Idealista and Fotocasa, Idealista was chosen for two reasons. Firstly, although it hasn't been around for that long it has more apartments and houses advertised. Secondly, and more important for this project, Idealista allows access to its API upon request as long as it is for research purposes which don't include monetary benefits.

With all the state of the art of both logical models and real estate there is still the relationship between the two to be researched which is the socio-economic background that can be found in Chapter 8. But finding that even if the real estate industry moves enormous quantities of money and affecting a countries economy as much as it does the technologies used are still antiquated (Klopp, 2016). Real estate companies are missing from innovation, for not looking at new data aggregations, analysis and distribution mechanisms. All this is aligned with the objectives of this project which summarized are to aggregate the data as automatized as possible, analysing the factors that impact the price of the real estate to create a logical model(s) to predict the prices.

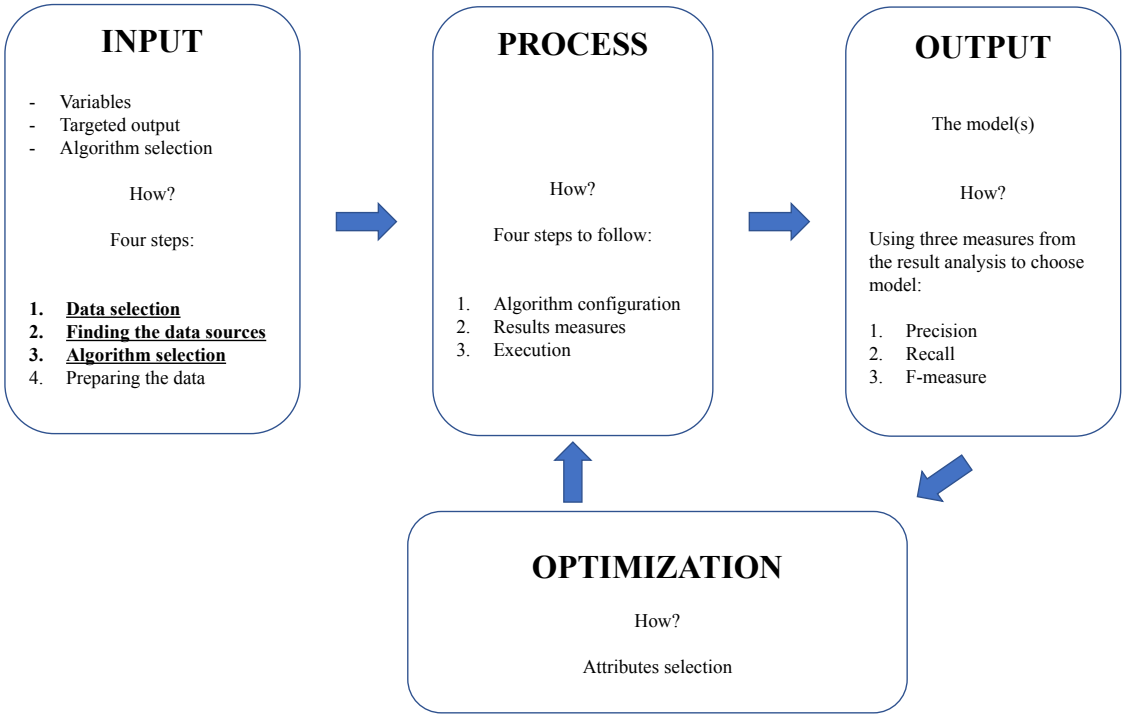
CHAPTER 4: ANALYSIS

During this chapter the analysis made in order to create the model will be described.

The problem this project is trying to solve is the classification of apartment and house prices using their characteristics. In order to achieve it, the steps to follow need to be identified, analysed and carried out.

This is a scheme of the overall process that the project will follow:

SCHEME II: PROJECT OVERALL SCHEME



So, looking at the previous graph this chapter will concentrate on the analysis made before getting into cleaning the data, steps one, two and three of the input. It mainly includes the research and analysis made to start up the project, it will leave out the technologies used as those will be analysed and discussed in Chapter 6.

4.1 Data selection

The variables chosen for this model can be assembled in four types of characteristics: surroundings, economy, region (country, city, district) and finally the characteristic of the apartment or house.

In order to show how each of these characteristics relate to the model there is a need to understand them and try to give each of them different values.

- Value of the land and the condition of the house/ apartment:

There are two factors to differentiate between, the land where it sits on and the actual value of the house. When talking about land we are referring to an appreciating asset (1), as its value will grow with time. This is quite simple to understand as with any limited asset the value will increase overtime as long as nothing out of the ordinary occurs. Like for example a natural disaster (tornado, earthquake) which will leave the area unusable. For the purpose of this project we are discarding the risk of natural disasters as we are looking at the Madrid, Spain area and not only its risk is negligible but as all de data from the model are from the same area it won't affect it.

On the other hand, the value of the apartment will decrease as time passes if there are no improvements and updates made. This will be taken into account by having a variable as to the state of the house.

When taking a look at the cadastre value there is something to consider, while it takes into account the first factor, the land value, the amount invested in the inside remodelling won't affect the land registry value.

The data to get the value of the land will be taken directly from the land registry, while the variables to see the value of the apartment related to the state of it will be taken from the Idealista website which let us know whether it's new, just remodelled, second-hand but good use or in need to be remodelled.

- Specific to the house/apartment:

These are the specifications of the house. Although there are many of them once the model is built and we take a look at which are significant probably most of them won't but the ones taken into account at the start are the following:

  - the square meters,
  - the orientation (south/east/west/north),
  - services such as lift, 24 h security and doorman
  - the antiquity of the building, house
  - number of rooms, bathrooms…

    - the floor where the flat is
    - whether it has parking space and if it is included in the price

- Surroundings:

Although it directly affects the land registry value, we still look at specific variables:

- Number of schools private and public.
- Public transport, bus stations, subway and train stations and lines.
- Parkland around.
- Hospitals near.
- Services such as restaurants, shops, businesses.
- Criminality.

- Region

When this project was first conceived, the scope was Madrid and its 21 districts but because of the complications that will be explained in this chapter the final decision was to analyse just one of them, Barrio de Salamanca.

The Barrio de Salamanca has six different neighbourhoods: Castellana, Lista, Guindalera, Fuente del Berro, Goya and Recoletos. The Barrio de Salamanca is considered one of the most expensive districts in Madrid but for each neighbourhood the price per m^2 changes as shown in the cadastral value.

Therefore, even if it will only predict for one district it will still have a regional value: the neighbourhood.

- Economy

The economy of the country or city clearly affects the value of real estate and if temporal data were available the IPV value could be used to use the model in the future.

The Idealista API doesn't give us data about when the advertisement was posted and therefore the economy was discarded as a variable. Still for future enhancements of this project the data of the IPV was collected.

Once the variables were identified, the next step is to find data sources that will allow as to get them.

4.2 Finding the data sources

The following table shows the name of the variable and the source of the data. The explanation of each data source can be found underneath the table.

TABLE I: VARIABLES AND THEIR DATA SOURCES

Variable	Width of street	District	Is it new?	Exterior/interior	Rooms	Bathrooms
Source	Google maps	Idealista	Idealista	Idealista	Idealista	Idealista

Variable	Floor	Neighbourhood	Size	Has lift	Status	Has parking space
Source	Idealista	Idealista	Idealista	Idealista	Idealista	Idealista

Variable	Is parking space included	Vs aprox	Vc aprox	Public middle schools	Private middle schools	Public high schools
Source	Idealista	Land registry	Land registry	Madrid government	Madrid government	Madrid government

Variable	Private high schools	Parkland	Hospitals	Restaurants	Social	Bus stations
Source	Madrid government	Google maps	Google maps	Google maps	Google maps	Google maps

Variable	Subway lines	Subway stations	Train stations	Train lines
Source	Google maps	Google maps	Google maps	Google maps

It is also worth mentioning the variables and data sources collected that finally were discarded for two reasons, either the fact that the model only used one district or that no time variable was found in the Idealista data. The following table shows the names of the variables and the data sources:

TABLE II: NOT USED VARIABLES AND THEIR DATA SOURCES

Variable	IPV	Criminal actions against people	Criminal actions against surroundings	Criminal actions with weapons	Criminal actions related to drugs
Source	Madrid government	Madrid government	Madrid government	Madrid government	Madrid government

Variable	Detainees	Accidents with people wounded	Accidents without people wounded	Alcohol consumed in public streets (+18)	Alcohol consumed in public streets by minors
Source	Madrid government	Madrid government	Madrid government	Madrid government	Madrid government

Variable	Population	Average price per district
Source	Madrid government	Madrid government

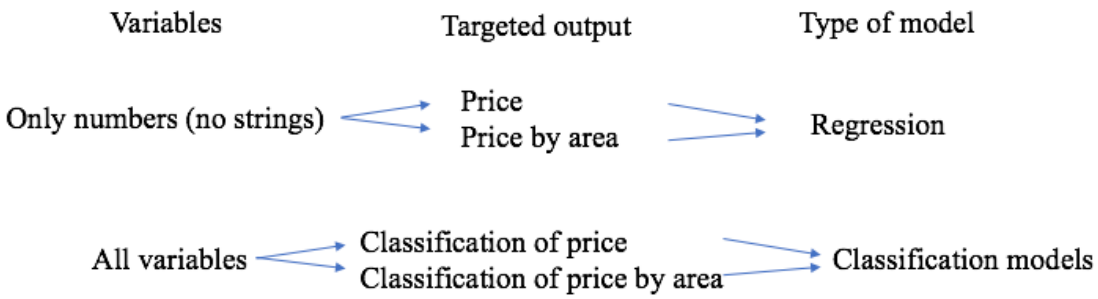
The automatization of the data collection depends on the source. While for Idealista there is the possibility of accessing the API with a code written either in R or in Python, access to the other variables was through a semi-automatized process. This will be further discussed in Chapter five.

4.3 Algorithm selection

In Chapter 3 after a first look at the different types of logical models was taken and because the database is made of labelled data it's clearly a supervised learning model.

For the targeted output, there were different options to consider, first if a regression model was used then there are two options: price and price by area. If a classification model was used then there will be endless options and a study of the distribution of our data has to be made. Also, once the distribution of data is clear the number of classes must be decided and depending on the length of time it takes for the models to be built there can be some additional tries to get the best model possible with as many classes as there can be. This is a scheme of the different possibilities:

SCHEME III: DIFFERENT MODEL POSSIBILITIES



Therefore, there will be 2 regression models and twice the number of the classification models there are available in WEKA times the number of different classifications that will be tried. Therefore, the number of models that will be compared to each other depends highly on the time they take and the time available.

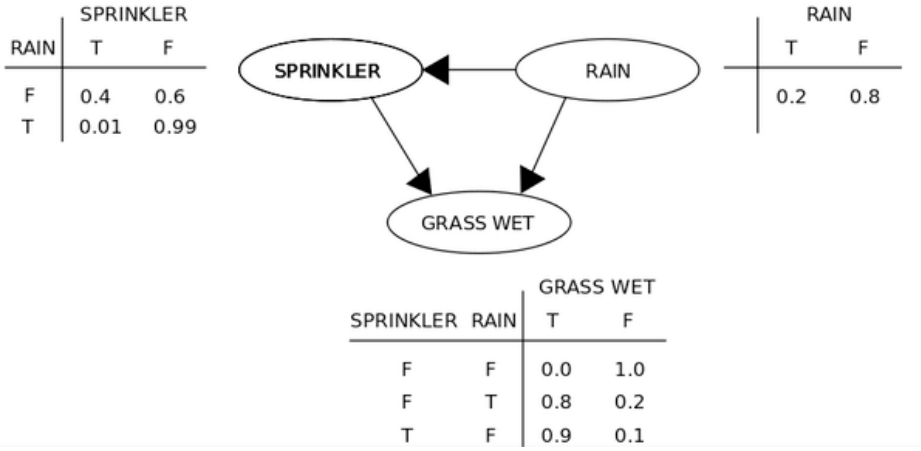
Further selection of algorithms will need the results of each type in order to see which has a higher precision. This is the list of the classifying algorithms:

- Bayes: BayesNet, NaiveBayes, NaiveBayes MultinominalText, NaiveBayes Updateable.
- Functions: Logistic, MultiLayer Perceptron, Simple Logistic, SMO
- Lazy: IBK, KSTAR, LWL.
- Meta: AdaBoostM1, AttrubuteSelectedClassifier, Bagging, ClassificationVia Regression, CVParameter Selection, Filtered Classifier, IterativeClassifier Optimizer, LogitBoost, Multi Class Classifier, Multi Class ClassifierUpdateable, Multischeme, Random Comitee, Randomizable Filtered Classifier, Random SubSpace, Stacking, Vote, Weighted Instances Handle Wrapper.
- Misic: Input Mapped Classifier.
- Rules: Decision Table, JRIP, OneR, Part, ZeroR.
- Trees: Decision Stump, Hoeffding Tree, J48, LMT, Random Forest, Random Tree, Rep Tree.

4.3.1 Bayes

The Bayes algorithms have an underlying probability model were the decisions of what instance belongs to which class is made by a probability calculation (StatSoft, 2013).

GRAPH II: BAYES EXAMPLE GRAPH (Wu, 2015)

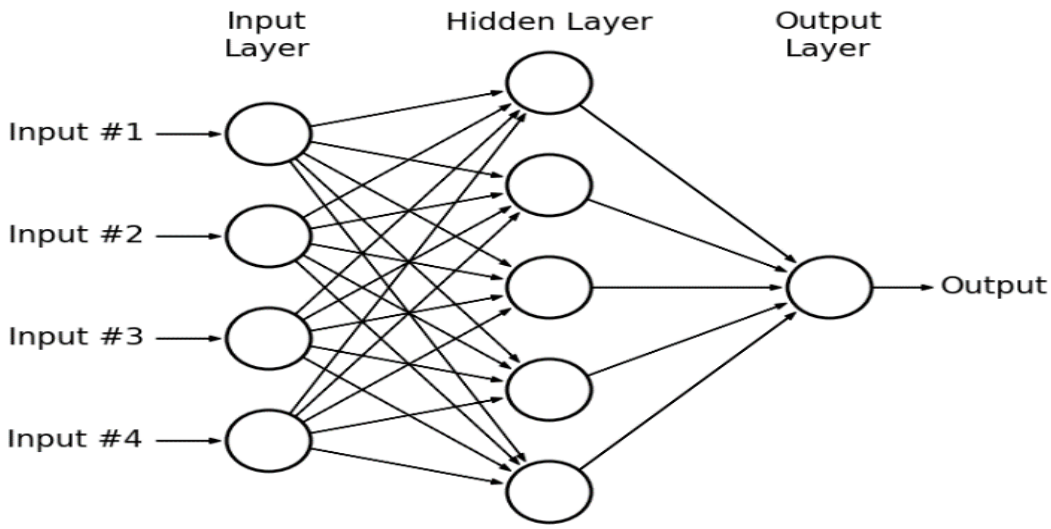


This graph is an example of a bayesanian network were the target is to know when the grass is wet and uses probability of it to solve the problem. For our project it won't be that simple as the variables are 27 and not 2.

4.3.2 Functions

They are used more commonly in regression models but the idea is to estimate a function. For example, multilayer perceptron is a classifier that uses backpropagation to classify instances. It creates a hidden layer to analyse the relationship between the input layer and the output layer as shown in the following graph (Negm, 2015).

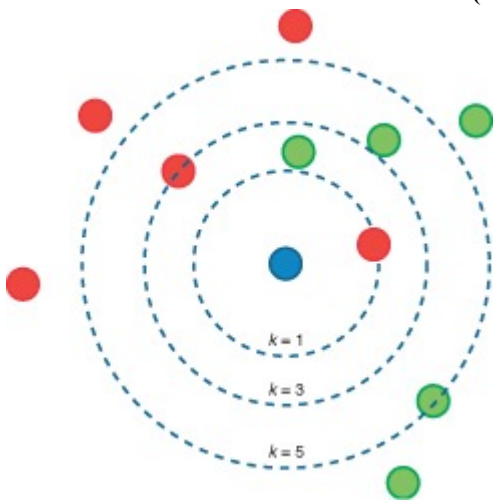
GRAPH III: MULTILAYER PERCEPTRON EXAMPLE NETWORK (Negm, 2015)



4.3.3 Lazy

The algorithms used in lazy learning are like k-Nearest Neighbours. The function is only approximated locally, they are among the simplest of all machine learning algorithms. It assigns weight to the neighbours so them are the ones that contribute more, instead of the distant ones (Morgun, 2017).

GRAPH IV: K-NEAREST NEIGHBOUR EXAMPLE (Morgun, 2017)

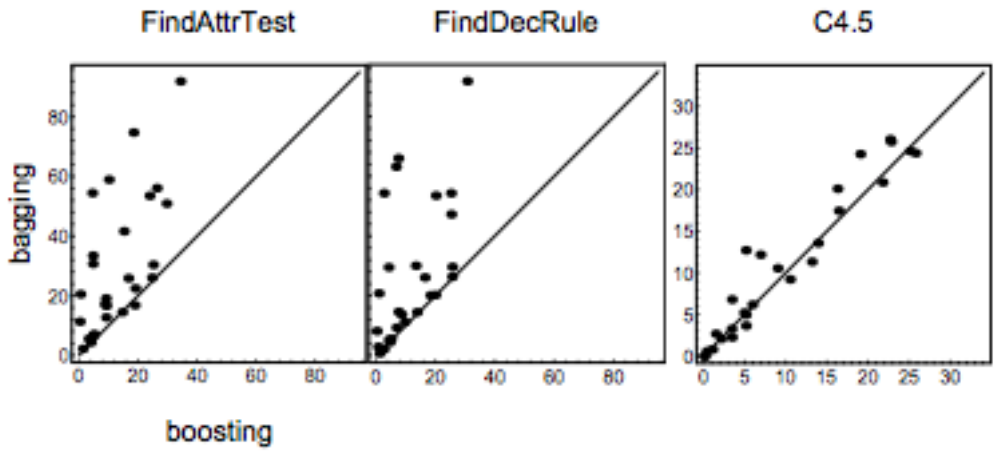


To understand the way these algorithms work the easiest thing is to see the Graph IV. If K is one then the blue dot will be predicted to be red, if K is three then it will be predicted to be red, if K is five then it will be green. It looks at who the neighbours are and it guesses that it is what it has more chances to be, what is more probable.

4.3.4 Meta

Meta are algorithms that combine or use multiple algorithms. They have the risk of overfitting which is a problem that will be explained in the next chapter. The main idea is boosting weak learners to make them stronger (Rusin, 2018).

GRAPH V: META EXAMPLES (Freund, 1996)



This graph is a comparison of boosting and bagging for different weak learners.

4.3.5 Misic

It is a wrapper classifier, it can work both with a new classifier or with one that has not seen before. It trains and tests data by building a mapping between the classifier that has been built and the incoming test instances' structure. If there is a variable that is not in the new instances it receives a missing value (Panthong, 2015).

4.3.6 Rules

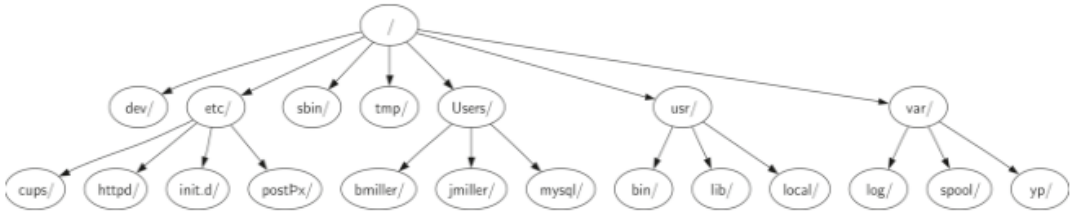
As the name shows these algorithms use rules. It will be very useful to our project as the ZeroR sets the baseline. So, for any machine learning algorithm to show that it adds value to the project it must show a higher accuracy. Another type for example is the Decision Table, whose output are a series of actions. They work like in programming if-then-else or switch-case (Borysowich, 2007). This is a simple example:

GRAPH VI: DECISION TABLE EXAMPLE (Borysowich, 2007)

Less than 50 Units Ordered	Y	Y	Y	Y	N	N	N	N
Cash on Delivery	Y	Y	N	N	Y	Y	N	N
Wholesale Outlet	Y	N	Y	N	Y	N	Y	N
Discount Rate 0%				X				
2%		X	X					X
4%	X					X	X	
6%					X			

4.3.7 Trees

GRAPH VII: TREE EXAMPLE (Patel, 2017)



This is a file system that has the structure of a tree, the same can be said about the html structure. For example the decision stump gives rules at every level and depending on the answer they follow one path or another (Patel, 2017).

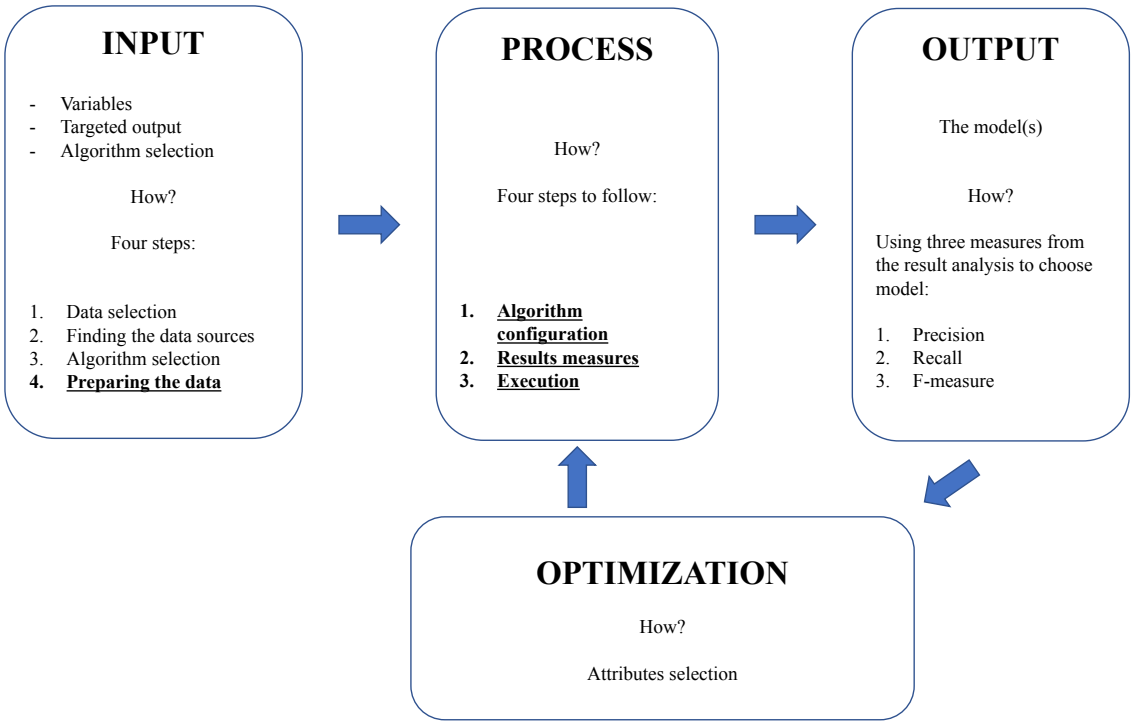
A part from these, regression algorithms will be used with the continuous target variable, the price without being classified.

In conclusion, once the model is created it must be compared to the Zero Rule to make sure the use of that type of algorithm has added value to the model. Furthermore, after the research into the algorithms, the next steps will not include to optimize it by boosting algorithms as the risk of overfitting is too great.

CHAPTER 5: DESIGN

The design of the project includes how the problem of classifying the houses and apartments prices was achieved which in our overall project scheme includes every step from accessing the data to getting the model. This chapter takes a in depth look at every step that had to be followed to get the different models.

SCHEME II: PROJECT OVERALL SCHEME



Looking at the scheme from the last chapter, it will go from the cleaning of the data to the optimization process, the results are recorded during these steps and can be found in the next chapter, Chapter 6. The technologies used and the automatization of the project will also be discussed during the next chapter, Chapter 6: Implementation as well as the final decision of the right model.

5.1 Preparing the data

This is the longest step of the project but once it has been achieved small changes are easy to make.

It can be divided into two sub-steps: cleaning the variables and finding a classification for the target variable.

5.1.1 Cleaning the data

This step mainly concentrated in two things, deleting things like a space before or at the end of a word and giving the others values:

- Width of the street: if at least one the roads that the building faces is either a two-way street or has more than one line then a 1 is given, if not a 0 is given.

- District, in this is case will always be Salamanca but for future purposes it would be a class value.

- If it is new, exterior, has lift, has parking space, is it included in price values is either 1 for yes or 0 for no (binary).

- Rooms, bathrooms, floor, size, Vs, Vc are numeric values that stay has they were given by idealista.

- Barrio is a string or class value.

- For public schools and private schools both middle and high schools there was no research found on how close they had to be to the home in order to be considered so a radius of 500 m was chosen as it involves a 5-10 minute walk at most which can be useful when talking about kids and the ability of them to walk themselves.

- The same 500 m radius was chosen for the parkland as it would mean a short walk.

- For hospitals, restaurants and leisure a classification was needed to assess how it would affect the price. If the hospital was between 0-1 km from the centre of the radius then it is considered close and the value of 1 was given, if it was between 2 and 5 km then it considered medium and a 2 was given and if it was further than that it is considered as having a hospital far and the value of 3 was given. The reason for choosing three classes instead of just 2 or 1 is that having a hospital too close may not be good as the noises from ambulances at night may decrease the price of a home.

- Bus lines were impossible to be considered as there are too many in the centre of Madrid and for all the model, doesn't matter how many districts are considered the Bus stations variable makes more sense to be considered as a binary value (either there are or there are not) because if there is a bus station its probable there is another not too far but if it is the same line it doesn't make mathematical sense to be considered twice.

- Subway stations and lines, unlike bus lines subway are much easier to be counted as there are not that many and the only question was how far they had to be to be considered as valuable. A 300 m was chosen as it will mean a 5 minute walk at most and when travelling inside Madrid more than that would mean it takes someone longer to get to the subway station than the subway ride.

- Train stations, although more than 300 m may not be too far the same argument as for the subway stations could be made and for the project to follow the same line of thought that radius was chosen. Train lines were counted as, like subway lines, they are not too many and their importance higher than bus lines.

5.1.2 Classification: price distribution

As it was mentioned before in order to us a classifying algorithm a distribution analysis had to be made. Two analysis were made, one for the price and the other for the price per square meter. Instead of choosing from the start the hypothesis was made that the models wouldn't take long to be built and therefore, not only the price and price per square meter was used but also both quartiles and deciles.

Once the data is organized then the next step will be to calculate the quartiles and deciles. The quartiles are calculated by dividing the data in four groups or classes:

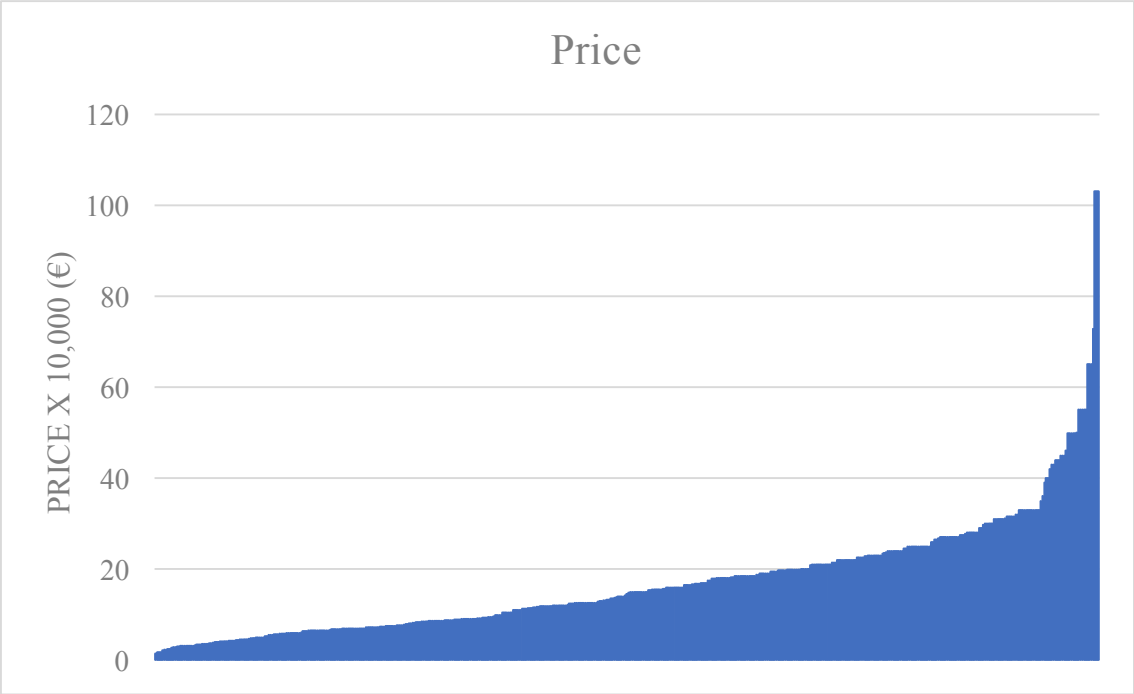
- Q_1 the lowest 25 % numbers.
- Q_2 the next lowest 25% numbers up to the median.
- Q_3 from the median to the highest 25% numbers.
- Q_4 the highest 25% numbers.

The deciles work the same way but instead of dividing the data in 25% (4 quartiles) we divide the data in 10% (10 deciles) getting D_1-D_{10}.

- Price

The price range of the data goes from 150,000 € to 10,300,000 € and this is how they are distributed.

GRAPH VIII: PRICE DISTRIBUTION



Seeing the graph, it's clear that for the first 495 prices there are no anomalies that weren't expected but once the price goes higher than 3,6 million there is not that much data. Still as the square meters of those houses are the highest ones, then there is no reason to do a further investigation on them as the following distribution will uncover if there is something really differential amongst them.

The values of the quartiles and deciles are in the following tables

TABLE III: QUARTILES (X10,000€)

Q_1	7,5
Q_2	14,5
Q_3	22,5
Q_4	103

TABLE IV: DECILES (X10,000€)

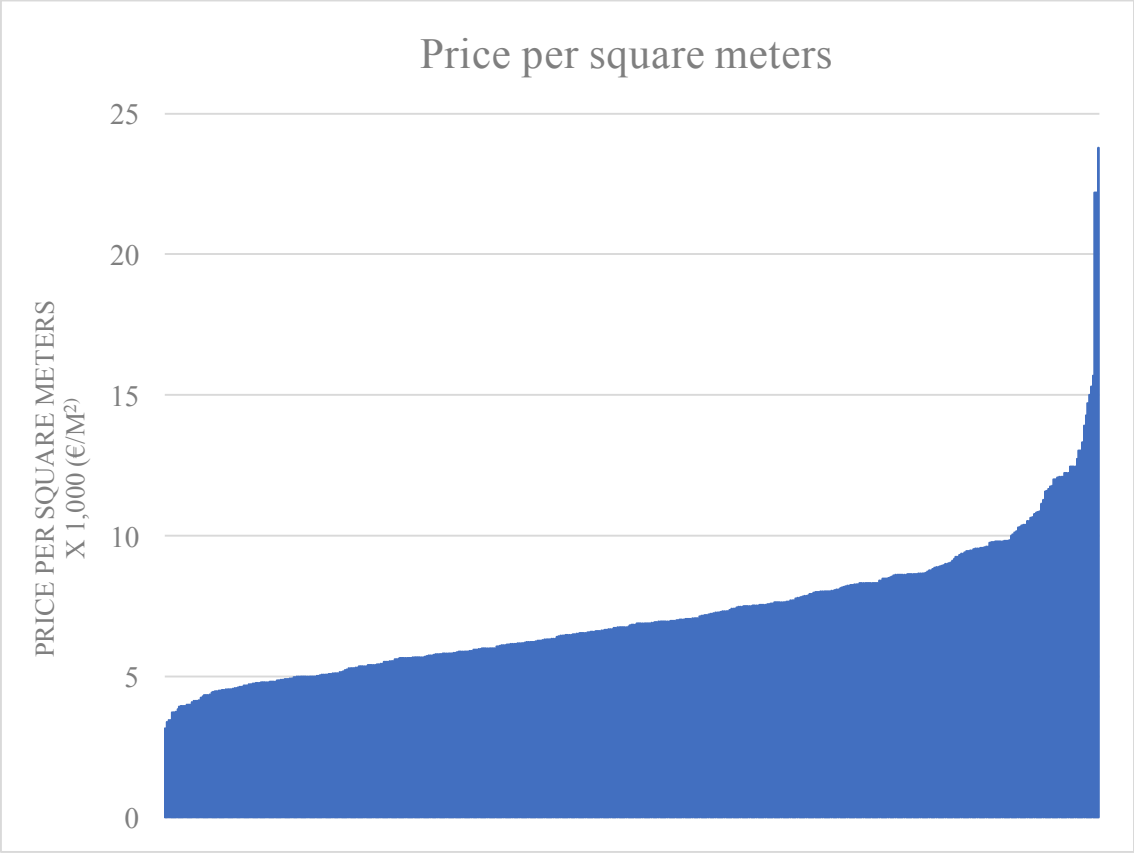
D_1	4,6
D_2	6,85
D_3	8,6
D_4	11,5
D_5	14,2
D_6	18

D_7	21
D_8	24,94
D_9	31
D_{10}	103

- Price by square meters

The price per square meters goes from 3158 €/m^2 to 23788 €/m^2 and this is how they are distributed:

GRAPH IX: PRICE PER SQUARE METER DISTRIBUTION



In this graph 518 prices per square meters seem to add up while there are still a few that seem to high. After a longer look was taken reasons for that prices appeared, such as how old the buildings were and the possibility of not being a flat but a house.

The values of the quartiles and deciles are in the following tables:

TABLE V: QUARTILES PER SQUARE METER (X1,000€/$m^{2)}$

Q_1	5,65
Q_2	6,85
Q_3	8,33
Q_4	23,78

TABLE VI: DECILES PER SQUARE METER (x1,000€/m^2)

D_1	4,77
D_2	5,29
D_3	5,82
D_4	6,25
D_5	6,82
D_6	7,32
D_7	8
D_8	8,65
D_9	9,82
D_{10}	23,79

Therefore, with the quartiles and deciles calculated the project is ready to create a csv and start trying different algorithms.

5.2 Algorithm configuration

As said before this project will both create a classifying model and a regression model in order to choose the best option. This sub-chapter will take a deeper look into the configuration of the algorithms that will be tried.

When creating a logical model there is never enough data that's why Cross Validation fold is needed. By using the training data if instead of Cross Validation, full the full training set was used then overfitting may appear.

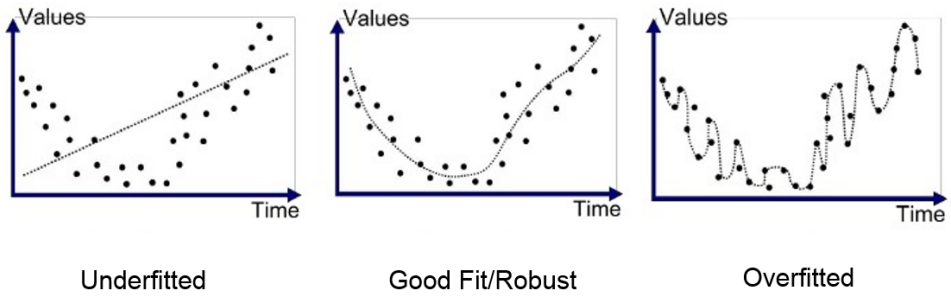
5.2.1 Overfitting vs underfitting

In machine learning overfitting happens when the model fits the data 'too' well. So much so, that is no longer useful for data that has never seen. For example, if in our data there weren't two apartments with the same size, overfitting would be if the model just decided to categorize depending on the exact size: if size=340 then class=Q4, if size=359 then class=Q3, if size=353 then class=Q2. Then in the example it may be able to correctly classify 100% of the data but it won't be able to correctly categorize an apartment that has never worked with before.

Underfitting is the exact opposite, for example trying to always predict by using a simple line: y=mx+n. It not only won't be able to predict but also it won't even be able to categorize the training data (Koehrsen, 2018).

The following image shows overfitting and underfitting quite clear:

IMAGE III: OVERFITTING VS UNDERFITTING (Koehrsen, 2018)



Although in the image II the differences are very easily seen, in any model with data a bit more complicated it is not that easy to realise when the project is over fitted. For that reason, if the algorithm gives a precision over 0.65 trying to boost the project may not be recommended.

5.2.2 Cross validation

In this sub-chapter, a description of cross validation and its folds will be given in order to use it and prevent overfitting.

Cross validation works the following way: It takes the data and divides it in the number of folds, it uses all the folds minus one as the training data. The remaining fold is used to validate the model. This process is repeated the number of folds and the final one is the average of all of the folds. Therefore, the overfitting phenomenon is prevented as none of the models created used all the data while the underfitting is prevented as in the final results all the data was used as the final result is the average of the previous models (Vanschoren, 2013).

5.3 Result measures

5.3.1. Result measures classification

For this project, the number of different models is quite high. WEKA has 40 different algorithms for classifying models and both quartiles and deciles will be tried, also both the whole price and the price per square meters will be used as targeted values. All this

means there will be 160 different models tried and before the optimization process begins there has to be a choice amongst the best models.
As for any engineering or economical project before a decision is taken there is a need to know which measures are the relevant ones. The measures are used to demonstrate whether or not the logical model is valid and how valid it is.

The four measures that will be recorded and from which the first decisions will be made are accuracy, precision, recall and F-measure. Another important measure is the time it takes for the model to be built. As there are only 525 apartments in our database, for now the time measure will be mostly ignored.

TABLE VII: SIMPLIFIED EXAMPLE OF POSSIBILITIES WHEN PREDICTING A CLASS (Koehrsen, 2018)

	Predicted class		
Actual class		Class= YES	Class= NO
	Class= YES	True positive	False negative
	Class= NO	False positive	True negative

These are the measures (Sunasra, 2017):

- Accuracy: The accuracy is simply out of all the instances, how many where correctly predicted. For this project: number of correctly classified apartments out of all of the apartments used (525). It will be looked at as a percentage.

- Precision: The precision is the ratio of correctly classified instances out of all the instances predicted of that class. In the table VII it will be one out of two of both classes, for yes it will be one true positive out of one true positive plus one false positive. For this project, the average of the precision will be used to analyse all the models and a further look per class will be analysed for the chosen models.

- Recall: The recall is the number of correctly classified instances out of all the actual instances of that class. In the example, it will be for the class yes: one true positive divided by one true positive plus one false negative. For the project, it will be in one class the number of correctly classified ones divided by all the ones in our database of that class.

- F-measure, also known as F1 score: is the weighted average of Precision and Recall. In an uneven class distribution is more useful than accuracy, but as the classification of this project has been done as an even distribution (for more details see sub-chapter 5.1.2) it will be approximately the same as accuracy.

5.3.2 Result measures regression

Although for regression there are also many algorithms that WEKA allows, this project will be concentrated in part of them.

The measures that will be looked at: correlation, mean absolute error, root mean squared error, relative absolute error and root relative squared error (BMJ, 2018):

- Correlation: is the relationship between the variables and the targeted value, it must always be between 0 and 1.

- Mean absolute error: is the measure of difference between the price and the predicted price.

- Root mean square error: is the standard deviation of the residuals.

- Relative absolute error: it a popular way of measuring the error rate but can only be used to compare models to each other. For this reason, this will be the measure to decide which models should be studied further.

- Root relative squared error: is the average of the total square error and decides it by the total squared error of the simple predictor.

5.4 Execution

Once the data is on an arff file, using Atom as it will be explained in Chapter 6, it is loaded in WEKA.

5.4.1 Execution of classifying algorithms

Then in the pre-process it allows for any variable to be dismissed but the first time the model is loaded the hypothesis is that all variables are useful. Weka then gives the option to choice the model type offering: Classify, Cluster, Associate. It also gives the option to Select Attributes, which will be used in the optimization process, and Visualize which once the model is chosen and the results gotten will be tried.

As the model is a Classifying model then the execution part of the project will mainly concentrate on that. In test options the cross-validation is chosen, as explained before, to avoid overfitting. Then in the classifier one by one all the allowed choices are tried and the result measures recorded, always taking especially care of the ZeroRule to know the model baseline.

This is done four times, one for each arff file: quartiles price complete, deciles price complete, quartiles price per square meter, deciles price per square meter.

5.4.2 Execution of regression algorithms

When using regression, the neighbourhood has to be taken out as no strings are allowed in this algorithm. Except from that all the execution process is the same. All the different algorithm used for the price per square meters will be used again for the complete price.

5.5 Optimization

For both regression and classification an analysis of the attributes is needed. In order to exclude those attributes that are not giving any value to the model the initial execution was required as the selection of attributes depends on the algorithm used.

So for the best result of model the selection of attributes was done by using the classifier attribute evaluation in WEKA using the algorithm with the best results.

For all the models the two attributes that didn't give the project any value were the bus lines and the restaurants. After a look, back to the database, was taken it was noticed that the values for both attributes were the same for all the data, therefore, making noise and not giving any differential value to the model.

Once both attributes were removed from the data step 5.4 was done again and the new values collected.

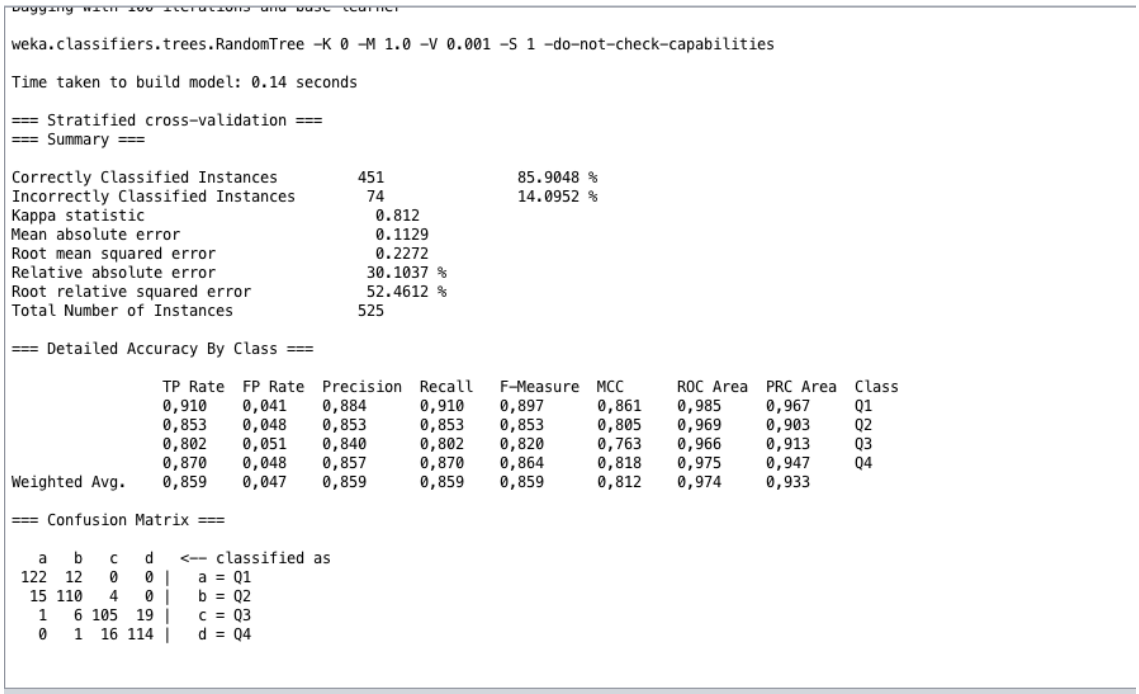
CHAPTER 6: IMPLEMENTATION

The implementation of this project is divided in two subchapters: the experiment implementation which includes the result analysis and the choice of model and the technologies implementation.

6.1 Experiment implementation

Once the steps described in Chapters 4 and 5 have been followed WEKA's classifier output deliver's a lot of information amongst which are the measures described in sub-chapter 5.3. This is an example of the classifier output:

IMAGE IV: WEKA OUTPUT EXAMPLE



```
Bagging with 100 iterations and base learner:

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.14 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         451               85.9048 %
Incorrectly Classified Instances        74               14.0952 %
Kappa statistic                          0.812
Mean absolute error                      0.1129
Root mean squared error                  0.2272
Relative absolute error                 30.1037 %
Root relative squared error             52.4612 %
Total Number of Instances              525

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,910    0,041    0,884      0,910    0,897      0,861    0,985     0,967     Q1
                 0,853    0,048    0,853      0,853    0,853      0,805    0,969     0,903     Q2
                 0,802    0,051    0,840      0,802    0,820      0,763    0,966     0,913     Q3
                 0,870    0,048    0,857      0,870    0,864      0,818    0,975     0,947     Q4
Weighted Avg.    0,859    0,047    0,859      0,859    0,859      0,812    0,974     0,933

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 122  12   0   0 |   a = Q1
  15 110   4   0 |   b = Q2
   1   6 105  19 |   c = Q3
   0   1  16 114 |   d = Q4
```

Using the accuracy of all the available models, before optimization, the following table was made:

TABLE VIII: ACCURACY PER MODEL PER OPTION

	Deciles/m^2	Quartiles/ m^2	Deciles	Quartiles
BAYESNET	22.48%	48.76%	48.95%	74.10%
NAIVEBAYES	26.29%	45.71%	48.57%	74.86%
NAVIEBAYESMULTINOMINALTEXT	10.67%	26.29%	10.86%	25.52%
NAIVEBAYESUPDETEABLE	26.29%	45.71%	48.57%	74.86%

LOGISTIC	30.29%	54.29%	53.9%	78.86%
MULTILAYEROERCEPTRON	32.76%	52.38%	57.71%	78.48%
SIMPLELOGISTIC	32.00%	52.00%	53.33%	79.81%
SMO	29.52%	48.38%	47.24%	77.90%
IBK	28.00%	49.95%	39.81%	64.57%
KSTAR	45.91%	57.52%	63.48%	75.62%
LWL	22.29%	43.05%	33.33%	59.62%
ADABOOSTM1	17.14%	40.76%	20.76%	49.90%
ATRIBUTESELECTEDCLASSIFIER	21.90%	54.86%	55.62%	78.48%
BAGGING	41.52%	56.95%	62.48%	82.10%
CLASSIFICATIONVIAREGRESSION	31.24%	56.19%	58.48%	81.33%
CVPARAMETERSELECTION	10.67%	26.29%	10.86%	25.52%
FILTEREDCLASSIFIER	20.00%	52.76%	57.33%	79.43%
ITERATIVEVLASSIFIEROPTIMIZER	32.76%	54.29%	56.38%	77.90%
LOGIT BOOST	32.76%	54.86%	56.38%	77.90%
MULTICLASSCLASSIFIER	30.10%	53.33%	47.43%	75.43%
MULTICLASSCLASSIFIERUPDETEABLE	13.71%	44.76%	25.14%	51.62%
MULTISCHEME	10.67%	26.29%	10.86%	25.52%
RANDOMCOMITEE	47.81%	65.33%	72.19%	85.52%
RANDOMIZABLEFILTEREDCLASSIFIER	41.14%	60.38%	60.95%	81.52%
RANDOMSUBSPACE	40.00%	59.43%	62.10%	80.76%
STACKING	10.67%	26.29%	10.86%	25.52%
VOTE	10.67%	26.29%	10.86%	25.52%
WEIGHTEDINSTANCESHANDLEWRAPPER	10.67%	26.29%	10.86%	25.52%
INPUTMAPPEDCLASSIFIER	10.67%	26.29%	10.86%	25.52%
DECISIONTABLE	18.29%	51.05%	47.43%	74.48%
JRIP	23.81%	55.81%	51.43%	77.14%
ONER	25.52%	43.62%	43.81%	66.29%
PART	36.76%	55.24%	59.24%	80.38%
ZEROR	10.67%	26.29%	10.86%	25.52%
DECISIONSTUMP	17.14%	40.76%	20.76%	49.90%
HOEFFDINGTREE	16.95%	37.9%	33.33%%	69.33%
J48	38.67%	57.52%	61.33%	84.19%
LMT	35.81%	55.81%	60.76%	78.10%
RANDOMFOREST	49.90%	66.10%	71.43%	84.95%
RANDOMTREE	44.00%	61.71%	63.43%	79.05%
REPTREE	33.71%	54.67%	56.00%	79.81%

Marked in green are the best accuracy results for each option, in blue is the baseline for all models and in red those models which didn't have a higher accuracy than the baseline and therefore had to be discarded as they didn't add value to the model.

Once the most recommended models were chosen the result measures described in subchapter 5.3.1 were recorded in the following table:

TABLE IX: PRECISION, RECALL AND F-MEASURE OF BEST PERFORMING MODELS

			Before optimization
RANDOMCOMITEE	**Deciles**	Precision	0,725
		Recall	0,722
		F-measure	0,722
RANDOMCOMITEE	**Quartiles**	Precision	0,855
		Recall	0,855
		F-measure	0,855
RANDOMFOREST	**Deciles/m^2**	Precision	0,494
		Recall	0,499
		F-measure	0,493
RANDOMFOREST	**Quartiles/m^2**	Precision	0,659
		Recall	0,661
		F-measure	0,659
RANDOMFOREST	**Deciles**	Precision	0,715
		Recall	0,714
		F-measure	0,714
RANDOMFOREST	**Quartiles**	Precision	0,849
		Recall	0,85
		F-measure	0,849

Although by know the data shows that it classifies much better the price than the price by square meters, the optimization process is done by all the models shown in Table IX.

As explained in the sub-chapter 5.5 Optimization, to better the result measures and therefore the model a selection of attributes is done using WEKA. To select attributes the evaluator chosen is the Classifier Attributes Eval and the method is Ranker.

Using this method, the attributes are ranked and given an average merit. The following image shows an example, particularly the Deciles/m^2 option, using the random forest algorithm. From that information, the attributes to be discarded were the Bus lines and the Restaurants:

IMAGE V: WEKA SELECTION OF ATTRIBUTES EXAMPLE

```
=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit      average rank  attribute
 0.279 +- 0.014      1   +- 0       14 Vcsaprox
 0.277 +- 0.014      2   +- 0       13 Vsaprox
 0.26  +- 0.017      3   +- 0        8 Size
 0.091 +- 0.006      4.9 +- 0.54    16 PrivateMiddleSchool
 0.087 +- 0.012      4.9 +- 0.94     6 Floor
 0.084 +- 0.007      5.2 +- 0.87    24 Subwaylines
 0.046 +- 0.006      8.5 +- 1.28    25 Subwaystations
 0.048 +- 0.013      8.7 +- 1.68    15 PublicMiddeleSchool
 0.045 +- 0.012      9.3 +- 2.53    18 PrivateHighSchool
 0.039 +- 0.008     10.2 +- 1.66    20 Health
 0.036 +- 0.007     10.8 +- 1.72     7 Neighbourhood
 0.031 +- 0.006     12.5 +- 2.16    11 HasParkingSpace
 0.028 +- 0.006     13.7 +- 2.05    12 Isincludedinprice
 0.028 +- 0.011     14.1 +- 3.56     5 Bathrooms
 0.026 +- 0.007     14.7 +- 2.33    17 PublicHighSchool
 0.026 +- 0.006     15.2 +- 2.48     3 Exterior
 0.017 +- 0.006     17.4 +- 2.2     27 Trainlines
 0.017 +- 0.006     17.9 +- 1.7     26 Trainstations
 0.016 +- 0.007     18.2 +- 2.4     10 Status
 0.015 +- 0.008     19.1 +- 2.39     1 Widthofstreet
 0.008 +- 0.005     20.8 +- 1.89    22 Social
 0.001 +- 0.003     23.7 +- 1.27     2 Wasitnew
-0     +- 0         24   +- 1.73    21 Restaurants
-0     +- 0         24.3 +- 1.27    23 Busstations
-0.001 +- 0.007     24.4 +- 2.65    19 Parkland
-0.006 +- 0.01      24.7 +- 2.45     4 Rooms
-0.001 +- 0.003     24.8 +- 1.33     9 HasLift
```

Once the attributes with a zero average are removed, the result measures were recorded and the selection of attributes done again for another optimization, then the process was done again choosing the values closest to zero until the models weren't optimizing anymore these were the result measures:

TABLE X: RESULTS AFTER OPTIMIZATION

			Before optimization	After 1st optimization only zero values	After 2nd optimization only <= 0,001 values	After 3rd optimization only closest to zero value
RANDOMCOMITEE	**Deciles**	Precision	0,725	0,705	**0,71**	0,71
		Recall	0,722	0,705	**0,712**	0,71
		F-measure	0,722	0,704	**0,71**	0,709
RANDOMCOMITEE	**Quartiles**	Precision	**0,855**	0,84	0,846	0,854
		Recall	**0,855**	0,84	0,846	0,853
		F-measure	**0,855**	0,84	0,846	0,853
RANDOMFOREST	**Deciles/ m^2**	Precision	**0,494**	0,488	0,487	0,486
		Recall	**0,499**	0,493	0,491	0,49
		F-measure	**0,493**	0,488	0,488	0,486
RANDOMFOREST	**Quartiles/ m^2**	Precision	0,659	**0,664**	0,656	0,648
		Recall	0,661	**0,665**	0,657	0,65
		F-measure	0,659	**0,664**	0,656	0,649
RANDOMFOREST	**Deciles**	Precision	0,715	0,717	**0,728**	0,713
		Recall	0,714	0,716	**0,726**	0,71
		F-measure	0,714	0,715	**0,724**	0,71
RANDOMFOREST	**Quartiles**	Precision	0,849	**0,859**	0,854	0,853
		Recall	0,85	**0,859**	0,853	0,853
		F-measure	0,849	**0,859**	0,853	0,853

Once the values in Table X are studied the best models are using the price without dividing it by the area and using the random forest algorithm. Whether choosing deciles or quartiles depends on how many classes are needed to predict and from now on they will be named as the Quartile model and the Decile model referring to the ones with the price NOT divided by the area.

By using the Quartile model there is an almost 86% chance of predicting the correct class while using the Decile model there is an almost 73% chance of predicting the correct one.

After the selection of attributes the ones discarded are the following

For Quartile model: Bus lines, Restaurants,

For Decile model: Bus lines, Restaurants and Parkland.

The question is what does the model predict when it doesn't do it right, in order to answer that question a look into the confusion matrix of both models was taken.

Confusion matrix of Quartile model:

a	b	c	d	<-- classified as
122	12	0	0	| a = Q1
15	110	4	0	| b = Q2
1	6	105	19	| c = Q3
0	1	16	114	| d = Q4

Confusion matrix of Decile model:

a	b	c	d	e	f	g	h	i	j	<-- classified as
43	9	0	0	0	0	0	0	0	0	| a = D1
8	33	6	3	2	0	0	0	0	0	| b = D2
1	12	36	6	2	0	0	0	0	0	| c = D3
0	4	4	37	2	1	1	0	0	0	| d = D4
0	1	7	2	39	2	0	0	1	0	| e = D5
0	1	0	3	4	39	5	1	4	1	| f = D6
0	0	0	0	0	8	38	6	3	1	| g = D7
0	0	0	1	0	3	8	23	6	2	| h = D8
0	0	0	0	0	1	1	2	46	3	| i = D9
0	0	0	0	0	0	1	1	4	47	| j = D10

Both confusion matrixes show a good prediction of quartiles and deciles as the highest numbers are always the ones that answer to the same predicted classification as actual classification and the second highest are the ones just before or after. For example, 46 of the Deciles 9 are predicted as Decile 9 and 3, the second highest number, of the Deciles 9 are predicted as Decile 10.

On the other hand, the regression models, although they weren't expected to achieve the same results they were still tried. These were the results of the result relative squared error without optimization:

TABLE XI: REGRESSION RESULTS BEFORE OPTIMIZATION

	Price/m^2	**Price**
GAUSSIAN PROCESSES	72.84%	38.54%
LINEAR REGRESION	73.9%	39.81%
MULTILAYERPERCEPTRON	97.86%	39.71%
SIMPLELINEARREGRESSION	87.49%	42.92%
SMOREG	71.95%	34.46%
IBK	79.90%	55.65%
KSTAR	73.63%	30.88%

LWL	83.04%	60.50%
ADDITVE REGRESSION	74.26%	44.35%
BAGGING	63.84%	30.25%
CVPARAMETERSELECTION	100.00%	100.00%
MULTISCHEME	100.00%	100.00%
RANDOMCOMITTEE	55.28%	22.82%
RANDOMIZABLEFILTEREDCLASSIFIER	74.14%	32.69%
RANDOMSUBSPACE	66.43%	32.49%
REGRESSIONBYDISCRETIZATION	63.34%	40.01%
STACKING	100.00%	100.00%
VOTE	100.00%	100.00%
WUGHTEDINSTANCESHANFLERWRAPPER	100.00%	100.00%
INPUTMAPPEDCLASSIFIER	100.00%	100.00%
DECISION TABLE	74.30%	40.66%
M5Rules	71.26%	34.70%
ZEROR	100.00%	100.00%
DECISION STUMP	85.31%	76.98%
M5P	67.10%	32.98%
RANDOMFOREST	55.53%	24.70%
RANDOMTREE	75.41%	31.92%
REPTREE	72.76%	34.98%

All regression models of the price per squared meters were ignored while the ones whose results were marked in green were considered for optimization as they were the ones that gave a 32% error or less. But after trying with all of them they don't get better results after optimization. Therefore, no optimization is needed for these models and the following table shows the results measures of those models with lowest relative squared error (the ones marked in green).

TABLE XII: RESULTS MEASURES REGRESSION

	Price				
	Correlation	**Mean absolute error**	**Root mean square error**	**Relative absolute error**	**Root relative square error**
KSTAR	0,8576	292061	713465	30,88%	53,49%
BAGGING	0,9122	286083	551191	30,25%	41,32%
RANDOMCOMITTEE	**0,9214**	**215817**	**519562**	**22,82%**	**38,95%**
RANDOMFOREST	**0,9267**	**233656**	**508718**	**24,70%**	**38,14%**
RANDOMTREE	0,8105	301926	785242	31,92%	58,87%

The chosen models from regression are the ones with the RandomComittee and RandomForest algorithms. Still, for the purpose of this project the classification models were studied more in depth.

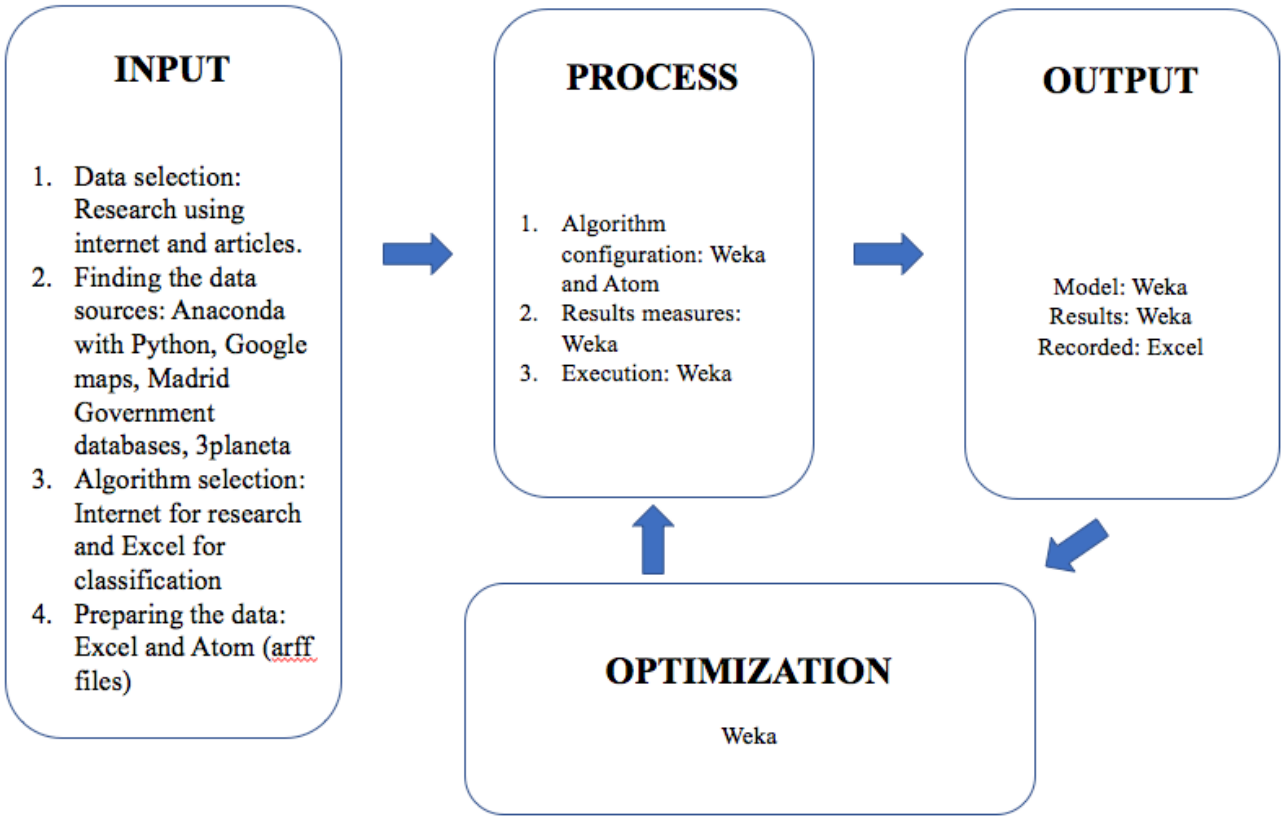
Therefore, there are four viable models, two using classification and two using regression.

6.2 Technologies implementation

During this project for each different step one or multiple technologies were used. For the length of this sub-chapter the different technologies will be discussed and the automatization degree of each step predicted and compared to what happened.

This is an overall view of the technologies:

SCHEME IV: OVERALL PROJECT TECHNOLOGIES



Because some technologies like Atom, Excel and WEKA are used in more than one step, instead of giving a description per technology there will be a description per technology per step.

6.2.1 Technologies: Step 1 data selection

The only in technologies used in this step were Google Chrome, Version 66.0.3359.181 (64 bits) and Adobe Acrobat Reader DC Version 2018.011.20038.

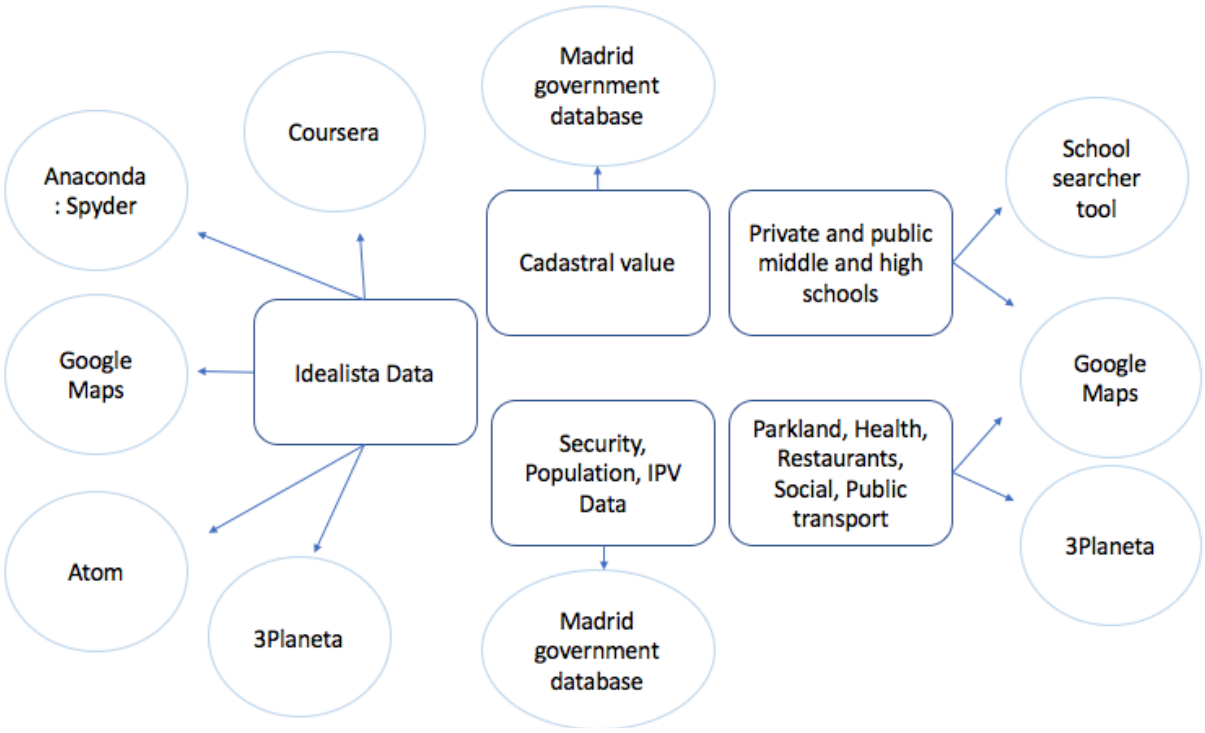
Both the automatization and the description of the use of these technologies are considered obvious and therefore no more explanation will be given.

6.2.2 Technologies: Step 2 finding the data sources

The technologies used were Anaconda: Spyder, Google Maps, Madrid Government databases, 3planeta.

As this step has diverse technologies for different functions the following scheme summarizes the different technologies used and for which data will be taken by or from them before describing them:

SCHEME V: OVERALL DATA SOURCES



Excel was where all the data ended up, therefore a description of the version is needed. Microsoft® Excel para Mac, Version 15.33 (170409)

The first data source used was Idealista and the technologies to use them were Anaconda: Spyder, in order to learn how to use python Coursera was used and research using Google Chrome. To get the data to Excel a sub-step to Atom was used.

Through Anaconda: Spyder using Python language the Idealista API was access the following code was the one used:

IMAGE VI: PYTHON CODE

```python
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
@author: anadelazaro
"""
#import base64
import httplib2
import requests
import requests.auth
import json



def get_token():
    client_auth = requests.auth.HTTPBasicAuth("6459q9c6u3dgruobblbtd1cec06fmqiu", "VzwYLCe6ZRQE")
    post_data = {"grant_type": "client_credentials",
                 "redirect_uri": "https://api.idealista.com/oauth/token"}
    response = requests.post("https://api.idealista.com/oauth/token",
                             auth=client_auth,
                             data=post_data)
    token_json = response.json()
    return token_json["access_token"]

def search_api(token):
    print ("Searching...")
    http_obj = httplib2.Http()
    lat_long = "40.430937,-3.681538"
    max_items = '150'
    distance = '212'
    propertyType='homes'
    operation = 'sale'

    url = "http://api.idealista.com/3.5/es/search?center="+lat_long+"&country=es&maxItems="+max_items+"&numPage=1&distance="
    +distance+"&propertyType="+propertyType+"&operation="+operation
    headers = {'Authorization' : 'Bearer ' + token}
    resp, content = http_obj.request(url,method='POST',headers=headers)
    return content

def extract_value(table, key):
    if key in table:
        return str(table[key])
    else:
        return ""

def load_json_file(name):
    print ("Loading file: "+name)
    try:
        fdata = open(name, 'r',encoding='unicode_escape')
        #with open(name) as json_data:
       #     d = json.load(json_data)
        fdata_content=fdata.read()
        fdata_content = fdata_content.encode("latin1", "strict")
        json_data = json.loads(fdata_content)
        fdata.close()
        return json_data
    except Exception as ex:
        print (ex)
        print ("Error reading file: "+name)
        return ""
```

```
60
61 def load_from_idealista():
62     try:
63         oauth_token=get_token()
64         results = search_api(oauth_token)
65         data = json.loads(results)
66         return data
67     except Exception as ex:
68         print (ex)
69         print ("Error reading from idealista")
70         return ""
71
72 def load(source):
73     if source == 1:
74         return load_json_file('salida_idealista.json')
75     if source == 2:
76         return load_json_file('idealistahomes1.json')
77     else:
78         return load_from_idealista()
79
80 if __name__ == "__main__":
81     data = load(3) #1 desde fichero o otro numero idealista load(2)
82     print (data)
83     for result in data['elementList']:
84
85         print ("Code:"+extract_value(result,'propertyCode')+
86                ";Thumbnail:"+extract_value(result,'thumbnail')+
87                ";Address:"+extract_value(result,'address')+
88                ";District:"+extract_value(result,'district')
89                +";New:"+extract_value(result,'newDevelopment')
90                +";Exterior:"+extract_value(result,'exterior')+
91                ";Rooms:"+extract_value(result,'rooms')+
92                ";Bathrooms:"+extract_value(result,'bathrooms')+
93                ";Floor:"+extract_value(result,'floor')+
94                ";Distance:"+extract_value(result,'distance')
95                +";Neighborhood:"+extract_value(result,'neighborhood')+
96                ";Price:"+extract_value(result,'price')+";PriceByArea:"
97                +extract_value(result,'priceByArea')+";Size:"+extract_value(result,'size')
98                +";Status:"+extract_value(result,'status')
99                +";HasLift:"+extract_value(result,'hasLift')
100               +";HasParkingSpace"+extract_value(result,'parkingSpace'))
101
```

The process is the following:

1st- A circle inside the district with radius being the distance parameter in the Code shown before was drawn in 3planeta.

2nd- From Google Maps by marking the same spot than the one used in 3Planeta's circle, the geographical coordinates are used.

3rd- In the Code in Spyder the new values are written down, the request made and then the json copied to Atom.

4th- The process is done again and again until all the District is covered.

5th- The data is imported to Excel.

Google Maps and 3Planeta's version is unknown as they are online tools.
The version of Spyder used was 3.2.4, and the language was Python3.
The version of Atom used was 1.25.1 x64

Once the request was made and received the json was copied to Atom from where after keeping only the elementlist it was imported into Excel.

The automatization of these steps is different for each. For the 3rd and 5th steps the process is mostly automatized while for the 1st and 2nd it is semi-automatized.

The second set of values taken were the Cadastral values. They were taken from the Madrid Government Webpage where a pdf of the approved values where found and written by hand in Excel. This process is mostly manual.

The third set of values taken were the Security, Population and IPV which were taken also from the Madrid Government Webpage but in this case the database was accessed and the data automatically downloaded to an Excel, therefore the process is completely automatic.

The fourth set of values were Parkland, Health, Restaurants, Social and Public transport were taken using both Google Maps and 3Planeta. The process is similar to the one used to get the geographical coordinates for the code in Spyder.

1st- The circle is drawn in the 3Planeta tool, then by using Google Maps the values are manually counted. The circle radius depends on the value being recorded, this radius is explained in Chapter 4: Analysis.

2nd- The data is recorded on Excel and the process is done again.

This process is semi-automatic, the online tools being automatic and the recording of the data manual.

The fifth set of values were the number of private, public, middle and high schools. The process is done using an open School searcher were by writing the coordinates and the radius every school inside that distance is described. The data is collected and recorded in Excel. This process is semi-automatic.

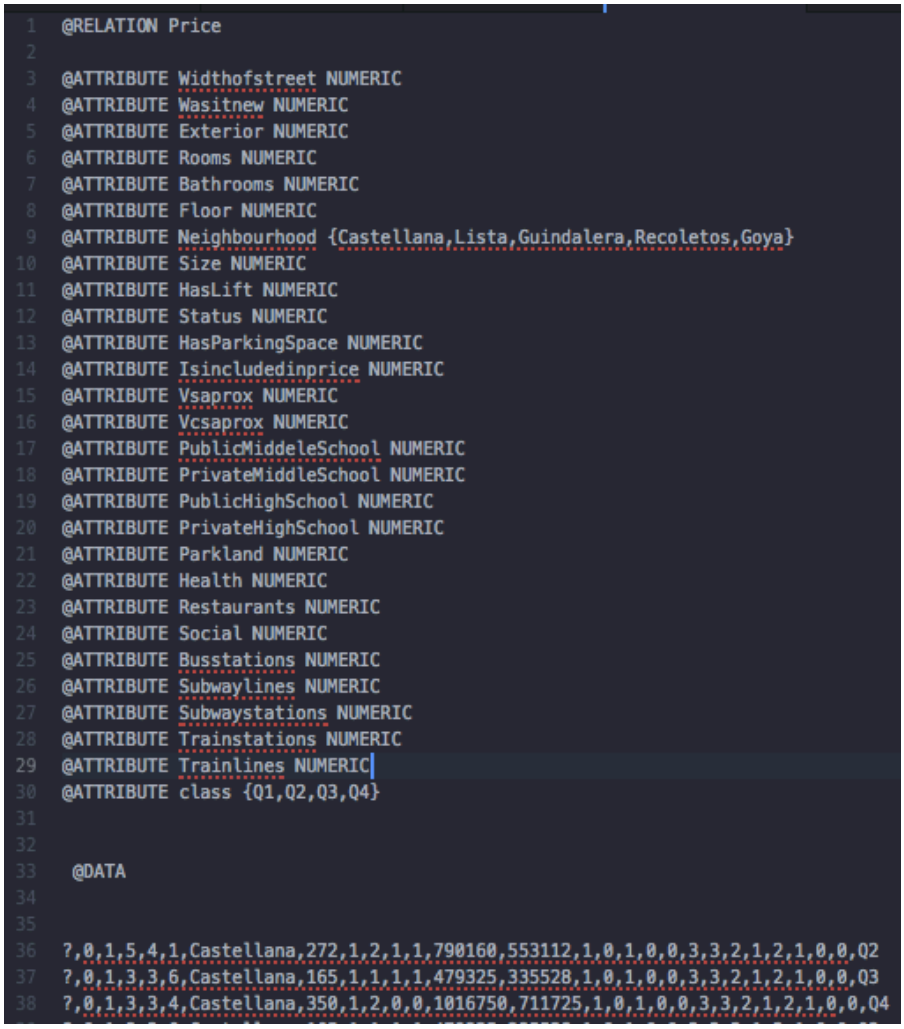
6.2.3 Technologies: Step 3 algorithm selection

This is done by researching using Google Chrome and by doing the distribution used for classification in Excel.

6.2.4 Technologies: Step 4 preparing the data

To prepare the data many Excel functions were needed and once the data was ready it was copied to Atom were spaces were replaced with ;. Then each attribute was typified. This is an example of one of the arff files in Atom.

IMAGE VII: ARFF FILE EXAMPLE



```
@RELATION Price

@ATTRIBUTE Widthofstreet NUMERIC
@ATTRIBUTE Wasitnew NUMERIC
@ATTRIBUTE Exterior NUMERIC
@ATTRIBUTE Rooms NUMERIC
@ATTRIBUTE Bathrooms NUMERIC
@ATTRIBUTE Floor NUMERIC
@ATTRIBUTE Neighbourhood {Castellana,Lista,Guindalera,Recoletos,Goya}
@ATTRIBUTE Size NUMERIC
@ATTRIBUTE HasLift NUMERIC
@ATTRIBUTE Status NUMERIC
@ATTRIBUTE HasParkingSpace NUMERIC
@ATTRIBUTE Isincludedinprice NUMERIC
@ATTRIBUTE Vsaprox NUMERIC
@ATTRIBUTE Vcsaprox NUMERIC
@ATTRIBUTE PublicMiddeleSchool NUMERIC
@ATTRIBUTE PrivateMiddleSchool NUMERIC
@ATTRIBUTE PublicHighSchool NUMERIC
@ATTRIBUTE PrivateHighSchool NUMERIC
@ATTRIBUTE Parkland NUMERIC
@ATTRIBUTE Health NUMERIC
@ATTRIBUTE Restaurants NUMERIC
@ATTRIBUTE Social NUMERIC
@ATTRIBUTE Busstations NUMERIC
@ATTRIBUTE Subwaylines NUMERIC
@ATTRIBUTE Subwaystations NUMERIC
@ATTRIBUTE Trainstations NUMERIC
@ATTRIBUTE Trainlines NUMERIC
@ATTRIBUTE class {Q1,Q2,Q3,Q4}


 @DATA


?,0,1,5,4,1,Castellana,272,1,2,1,1,790160,553112,1,0,1,0,0,3,3,2,1,2,1,0,0,Q2
?,0,1,3,3,6,Castellana,165,1,1,1,1,479325,335528,1,0,1,0,0,3,3,2,1,2,1,0,0,Q3
?,0,1,3,3,4,Castellana,350,1,2,0,0,1016750,711725,1,0,1,0,0,3,3,2,1,2,1,0,0,Q4
```

This process can be defined as semi-automatic because even if some of the preparing od the data was manually the use of Excel functions can be considered automatic, specially the use of Quartile and Decile functions.

6.2.4 Technologies: WEKA

For the rest of the project everything was done using WEKA for more details on each step go to Chapter 5: Design as all of them were explained there. Every process done via WEKA is automatic although some of the configurations were manually set, the most important part was done by the WEKA platform.

CHAPTER 7: PLANNING AND FINANCES

In this chapter, both the different costs of the project and the planning of it will be described in detail.

7.1 Planning

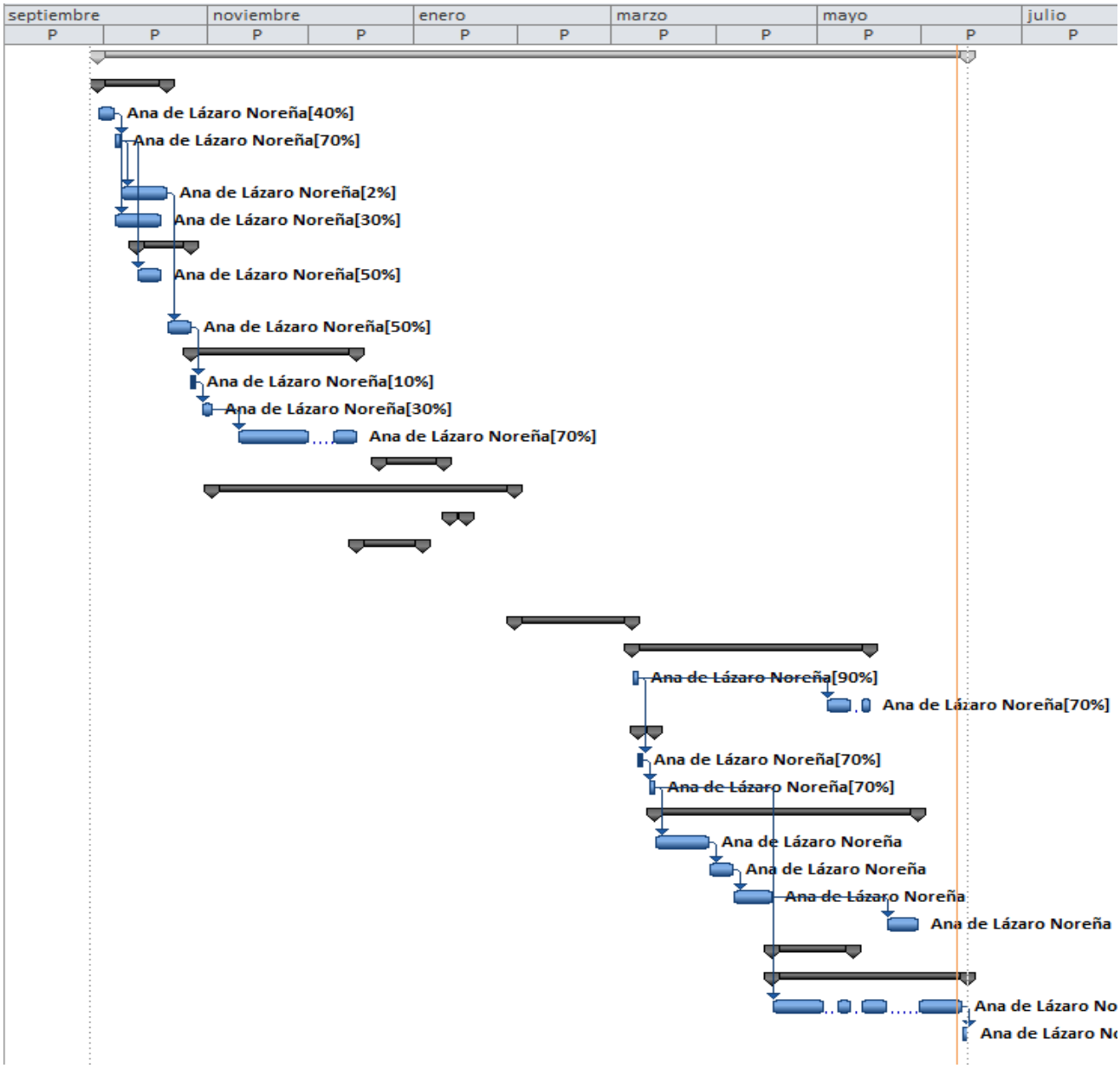
These are the steps that were followed to be able to do the project, as it can be seen they are structured as the project was with some additions.

- Task 1: Kick-starting the project:
  - Task 1.1: Describing the model objectives.
  - Task 1.2: Finding an API access to house prices.
  - Task 1.3: Request and answer.
  - Task 1.4: State of the art.
- Task 2: Learning Python:
  - Task 2.1: Coursera: 'Using Python to Access Web Data'.
  - Task 2.2: Online research.
- Task 3: Accessing Idealista's API:
  - Task 3.1: Installing Anaconda.
  - Task 3.2: Creating the code.
  - Task 3.3: Making the requests.
- Task 4: Transferring data to Excel:
  - Task 4.1: From Spyder to Atom.
  - Task 4.2: From Atom to Excel.
- Task 5: Cadastral value:
  - Task 5.1: Research.
  - Task 5.2: Finding the values.
  - Task 5.3: Getting the data to Excel.
  - Task 5.4: Hypothesis and Calculation.

- Task 6: Security data, IPV:
  - Task 6.1: Finding the source.
  - Task 6.2: Getting the data to Excel.
- Task 7: Population, Parkland, Health, Restaurants, Social, Public Transport:
  - Task 7.1: Finding the source.
  - Task 7.2: Getting the data to Excel.
- Task 8: Cleaning the data
- Task 9: Classification:
  - Task 9.1: Research.
  - Task 9.2: Distribution.
- Task 10: All data from Excel to Atom:
  - Task 10.1: New clean Excel.
  - Task 10.2: Excel to Atom.
- Task 11: WEKA, classes:
  - Task 11.1: Formation on WEKA.
  - Task 11.2: Execution.
  - Task 11.3: Research on optimization.
  - Task 11.4: Optimization.
- Task 12: WEKA, regression:
  - Task 12.1: Execution.
  - Task 12.2: Optimization.
- Task 13: Project documentation:
  - Task 13.1: Dissertation.
  - Task 13.2: Presentation.

This is the diagram part of the grant graph, for more information please go to the Appendix II:

GRAPH X: GRANT GRAPH



7.2 Finances

During this sub-chapter, a look will be taken into the costs of the project, for the purpose of giving an approximate budget to anyone who wants to acquire the model.

It is important to mention that a hypothesis was made that Idealista gave their permission for there to be profit, as it can be found in sub-chapter 8.2 in Idealista's Terms and Conditions (Idealista, 2018).

The costs will be divided in three sections:

- Labour Costs
- Material Costs
- Indirect Costs

7.2.1 Labour costs

During the planning, the labour hours were recorded. The following table shows the number of hours that were dedicated to the project:

TABLE XIII: HOURS PER TASK

Summary task	Task	Hours
Kick-starting the project	Describing the model	9.6 h
	Finding an API access to house prices.	11.2 h
	Request and answer	1.6 h
	State of the art	24 h
Learning Python	Coursera: 'Using python to access Web Data'	20 h
	Online research	20 h
Accessing Idealista's API	Installing Anaconda	0.8 h
	Creating the code	7,2 h
	Making the requests	112 h
Transferring data to Excel	From Spyder to Atom	16 h
	From Atom to Excel	24 h
Cadastral value	Research	33.6 h
	Finding the values	28 h
	Getting the data to Excel	19.2 h
	Hypothesis and Calculation	11.2 h
Security data, IPV	Finding the source	11.2 h
	Getting the data to Excel	5,6 h
Population, Parkland, Health, Restaurants, Social, Public transport	Finding the source	28 h
	Getting the data to Excel	28 h
Cleaning the data	Cleaning the data	140 h
Classification	Research	14.4 h
	Distribution	44.8 h
All data from Excel To Atom	New 'clean' Excel	5.6 h
	Excel to Atom	11.2 h
WEKA, classes	Formation on WEKA	86 h
	Execution	40 h
	Research on optimization	64 h
	Optimization	56 h
WEKA, regression	Execution	51.2 h
	Optimization	32 h
Project documentation	Dissertation	48 h

	Presentation	3.2 h
Total		**1007.6 h**

Taking that the cost for an Engineer is approximately 20€/h, then the Labour cost will be 1007.6 h x 20 € / h = 20,152 €.

7.2.2 Material costs

The following table includes all the costs from the material needed to carry out the project. The cost for all equipment is a 25% of their total cost as they were used for almost a year.

TABLE XIV: MATERIALS AND THEIR COST

Material	**Total cost**	**Cost to the project**
Computer (Laptop Macbook air)	1,105.59 €	276.40 €
External Keyboard	149.00 €	37.25 €
Extra Screen	119.99 €	30.00 €
Extra Mouse	14.99 €	3.75 €
HDMI cable to computer	10.99 €	2.75 €
Total		**350.15 €**

7.2.3 Indirect costs

The indirect cost will be calculated as an 18% of the direct costs which includes both the labour costs and the material costs. Indirect cost = (20,152+350.15)x0.18= 3690.39€

7.2.4 Total costs

TABLE XV: TOTAL COST

Type of cost	**Quantity**
Labour Cost	20,152 €
Material Cost	350.15 €
Indirect Cost	3,690.39 €
Benefit (10% total cost)	2,419.25 €
IVA (21%)	5,588.48 €
Total	**32,200.27€**

CHAPTER 8: LEGAL AND SOCIO-ECONOMIC BACKGROUND

8.1 Legal implications for the cadastral value

The Royal Legislative Degree 1/2004 of March 5th says that the cadastre description must include the physical, economic and juridical characteristics, amongst which the localization and the cadastre reference, the area, the use or the purpose, the building quality, graphic representation, cadastre value and ownership must be included.

On the other hand, the Royal Legislative Degree 2/2011 of March 11th adds that the tax identification or the foreign identity number, if needed must be included in the cadastre register. But more importantly for the purpose of this project it adds that the description and graph of the characteristics will be incorporated to documents with public access (Delgado Ramos, 2011).

In 2003, the Government created the Cadastral Electronic Site which objective was to provide access to other Administrations. It works as a data bank containing all the information of real estate in Spain (physical attributes, cadastral value, ownership).

The complete access is only available to the owners of each particular property and Governmental Administrations as one of its main uses is taxation. Both Spanish laws and European union consider the economic and ownership as protected data (Gobierno de España).

Access has been granted to private companies such as BBVA which connect to the cadastre value of a house to create the model to show the approximate value of any house. Still that access is only for the economic information and never for the ownership.

Ownership is, for example, used for scholarships. Institutions have the ability to know how much real estate the parent of any student asking for a scholarship owns and how much is worth.

8.2 Legal implications for the data from Idealista

Since May 25th 2018 a new General Data Protection Regulation, from now on referred to as GDPR, has been implemented. As this project works with data uploaded by users, although the access was made before May 25th, this chapter will take a look at the law and how it affects Idealista.

The main objective of GDPR is to unify the EU in their data protection giving the citizens and residents control over their personal data. The key points it contains are the following (Digital Guardian, 2018):

- Requiring the consent of the user for data processing.

- It makes the companies to anonymize the data in order to protect privacy.

- Companies have to let now the users if and when there is a data breach.

- If the company handles private information there needs to be appointed a data officer to make sure GDPR is being complied with.

- It affects to all EU citizens' personal data even if the company is not EU based.

As Idealista collects the IP address, the type of device and the localization therefore it must comply with the new GDPR law but as our project doesn't have any information about the users of Idealista, except the part they shared to everyone else the law doesn't have any repercussions to the project.

What does affect it is what Idealista states in their privacy policy as in order to access the data this project acceptance is compulsory.

Idealista states the following in their privacy policy and general conditions (Idealista, 2017).

No user is allowed:

- To publish content that is racist, xenophobic, obscene, derogatory or that incites, involves, or promotes criminal, violent or defamatory acts on the grounds of age, gender, religion, beliefs, or that involves the defence of terrorism.
- To be involved with acts that infringes the rights of others in any way.

- To use the Website and Apps for illicit or commercial purposes, for the purpose of making a profit, or for the resale, transfer or disposal for the use or benefit of any other person or entity.

- Cause damage to IT systems owned by Idealista or third parties, or introduce or spread computer viruses or anything else that could damage the IT systems.

- To carry out acts infringing on Idealista's Intellectual and Industrial Property rights, or the rights of third parties.

- To carry out acts that in any way jeopardise or that could jeopardise the reputation of Idealista or third parties.

- To use the identity and/or access keys or passwords belonging to another Registered User without their consent.

- To republish images or publish images with any kinds of markers, watermarks, logos or text.

- Publish images that are not relevant to the listing.

- Make any false or fraudulent contact.

- Access, control or copy any content or information included on the Website and Apps using any kind of robot, spider, scraper or any other automatic or manual process to do so for any purpose, without their express written permission.

- Violate the restrictions contained in any notice on the exclusion of robots included on the Website, or bypass or circumvent other measures used to prevent or limit access to the Website.

- Take any action that imposes or that may impose an unreasonable or disproportionately large load to their infrastructure.

- Copy, display or otherwise incorporate any content from the Website and Apps into any other website without their prior written permission.

- Use the Website and Apps and/or Services and Additional Services contrary to these General Conditions and/or the Specific Conditions, where relevant.

- To carry out any of the above-mentioned actions, the user will be liable for any damages arising from such direct or indirect breach of these General Conditions, and the user agrees to hold Idealista harmless. In the event that these General Conditions or the Specific Conditions are breached, they reserve the right to unilaterally cancel, at their sole discretion and without prior notice, any user's access to, use and/or registration for their Website and Apps and Services, without this in any way giving rise to any form of compensation.

In conclusion, for the purpose of this project without written permission the information supplied here and the uses of our model may never be connected to any profit. Also, the number of requests using the API can't overload their servers, although the following measures were taken by Idealista to prevent that: It limited the free access to 100req/month and limiting them by 1 req/sec.

8.3 Socio-economic background

The real estate market although always considered a conservative market in the last few years it has been shaken up by digitalization. It first began to be palpable in real estate marketing with the appearance of housing portals.

Due to all the characteristics shown of each apartment in these housing portals, 3D tours, photographs, comments of previous owners or renters, the customers' expectations are much harder to meet. Before the Technology era exploded in order to sell your home the most important factor was to find the right agency which also implied high commissions. Nowadays it is more useful to advertise in housing portals or even hiring a social networks expert. Even real estate agencies do so; even before construction has begun the upcoming apartments are advertised on them.

Also, related to this, digitalization in the real estate has created investment opportunities. Investing on real estate has always been only available to the wealthy but with crowdfunding platforms it has allowed for the medium workers to pool their money together and be able to buy investment opportunities. Although a couple decades ago it was said that real estate is a 'sure thing' as the value only grows since the 2006 crisis it has been proven false, still it is one of the safest investment opportunities and these crowdfunding platforms has made them available to almost all.
Another repercussion of the crowdfunding platforms is that they have taken a bit of power away from banks, it use to be that only them could lone their customers the money needed to buy a house and the only competition they found was amongst other banks. Not anymore, they now need to keep their prices competitive with crowdfunding platforms such as Patchofland.com.

Another important aspect is the need for the real estate agencies not only to create these kinds of models but to update them constantly and as the technologies improve to keep investing on research and innovation to keep up with the market (Klopp 2016). It is commonly believed that technology, big data and the rise of "space as a service" will disrupt the traditional property valuation model and it is now, if not ten year ago, when the real estate agencies need to improve their data collection, analytics and digitalization (PricewaterhouseCooper, 2017).

This project gives a small insight on all that can be achieved in the real estate business with data collection and analysis.

CHAPTER 9: CONCLUSIONS AND FUTURE WORK

9.1 Conclusions

As it was mentioned from the start the Real Estate business is not an easily predicted one. It fluctuates sometimes for none apparent reasons and similar houses have different prices without an explanation to be found.

Still this project is considered successful as it delivered not one but four models with their own strengths that can predict what can be considered a good prediction. For the models with classes an 86% precision was achieved for Quartiles and 73% for Deciles, but not only that but if a deeper look is taken to their confusion matrixes it can be seen that when the prediction was wrong it usually predicted the next closest class.

This conclusion won't select a final model as the best or more accurate one as each has their different functions. The regression models will give their users an exact price, with the error to be considered while the classifying although more accurate will only tell them a range of values.

Also, choosing between the Quartile model or the Decile model won't be done as while the Quartile has an excellent precision their class ranges are much bigger (approximately 2.5 times the decile ones) than the Decile ones.

Finally, to sum up, this project shows that a logical model with good results is possible for the Real Estate market and furthermore it delivers a learn how to roadmap to logical models using platforms free and available to anyone who wishes to use them.

9.2 Future work

This project is after all an end of studies project and has to be done at the same time as finishing other courses but with more time a lot of enhancements could have been done. In the following two sub-chapters some of them and a short analysis made are described.

9.2.1 Data enhancements

As it was discussed all through the project there was the limit of time and resources which were an impediment to creating a model for all of Madrid and that will take time, and therefore the economy into consideration. So, for future enhancements the data of the economy and district was collected and therefore, there wouldn't be a need to start from scratch. These would be the added variables:

- Security: the Madrid Government records different types of criminal offences that happen all around Madrid but the data is collected in a district level and as there was only one district analysed it can't be used until more are used. This will also open up the opportunity to analyse how much the criminality affects the price and which criminal offence has a higher impact on it.

- District: it will allow to see differences of price depending on the district.

- Population: as it is recorded by district it wasn't used during the project but it would be quite interesting to see how the price behaves when the population is high or low.

- IPV: that is the dwelling value factor. The Spanish Government calculates every term this value and if the time when the price of the home was uploaded were recorded it would allow to have a model that works in the future and not only at the moment. Also, the time factor its believed to allow for a model with better percentages of recall.

Some changes would have to be added in the values of public transport as before the district was in the city centre then the importance of a station being near may not be 300 meters but a much higher value. For example, if we are talking about Tres Cantos having a station 600 m away may be considered as having one close while in the Barrio de Salamanca there will always be one nearer and that one will be considered, as it has been for this project, too far. The same can be said for hospitals as having one just in a 5 minute drive can be considered as close while not the same can be said for Madrid centre.

Furthermore, the Government of Madrid has an average price per district per area and if more than one district is used maybe an analysis of whether using this price would work the same way inside the model than using some of the regional and localization values.

This model was done with only 525 apartments and 27 variables, but as more data is added that would implied the need for computers with higher process capabilities and more time spent cleaning the data, which will mean a higher cost of the project.

Also, if databases with some of the variables that had to be found in google maps, such as parkland, hospitals, width of street and bus, train and subway stations and lines it will cut the time of getting the data by almost 90%.

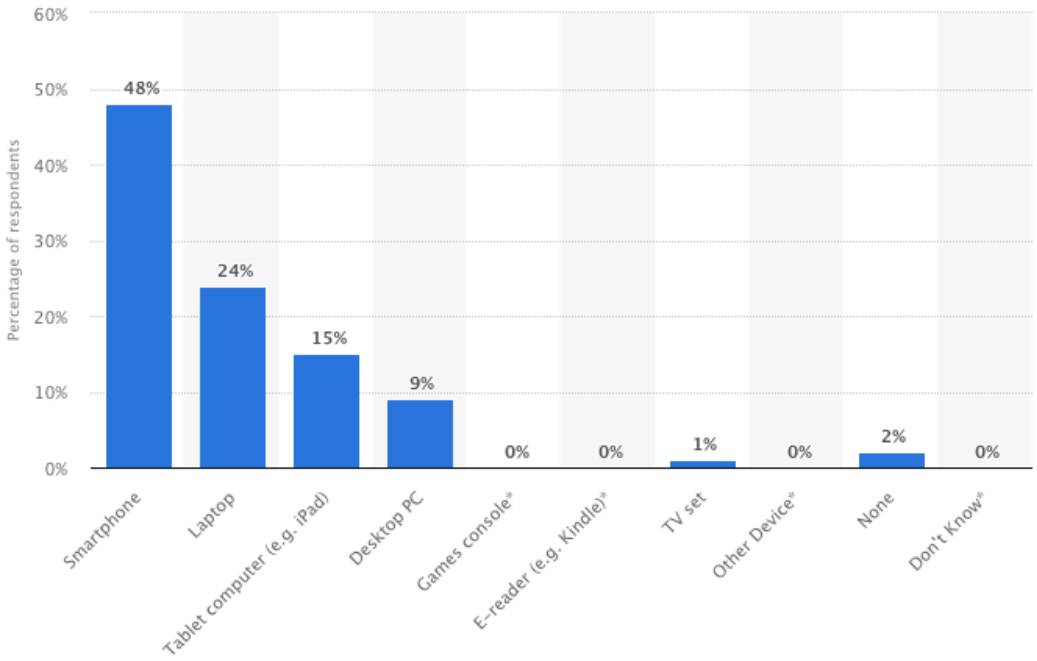
Finally, as mentioned before not having an exact address of most of the apartments forced most of the locational values to be approximate. By getting an API access to other housing portals with as much information as Idealista but with addresses it could make the model better as the data will be less incomplete. Also, it would allow our approximation of the cadastral value to be more exact.

9.2.2 Other enhancements

A web page may be useful to deliver this information for future users and using idealista price both could be shown in order to recommend to the user to try to get the price down or if it is reasonable.

A statistic made this by the UK showed that 48% of people who were asked what is the most important device they use to connect to the internet was a smartphone. For this reason, the presentation of the model and their used could be improved by adding an iOS or Android programmer who will be able to create an app, making the prediction of price easier for the consumer (Statista, 2018).

GRAPH XI: STATISTICS ON DEVICES USAGE GRAPH (Statista, 2018)



BIBLIOGRAPHY

This project was referenced following the APA format. These are the references used:

BMJ. (2018). 11. Correlation and regression. Retrieved June 2, 2018, from https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression

Borysowich, C. (2007). Building Decision Tables. Retrieved May 10, 2018, from https://it.toolbox.com/blogs/craigborysowich/building-decision-tables-042507

Certicalia. (n.d.). Calcular valor catastral. Retrieved June 5, 2018, from https://www.certicalia.com/calcular-valor-catastral

DCC, A. (2017, October 05). Catastro. Retrieved from http://elcatastro.blogspot.com/2011/04/ejemplo-de-calculo-de-valor-catastral_07.html

DCC, A. (2017, October 05). Cálculo de Valor de suelo urbano por repercusión. Retrieved from http://elcatastro.blogspot.com/2011/04/ejemplo-de-calculo-del-valor-del-suelo.html

DCC, A. (2017, October 05). Coeficientes correctores del suelo. Retrieved from http://elcatastro.blogspot.com/2011/04/coeficientes-correctores-del-suelo.html

Delgado Ramos, J. (2011, April 26). EFECTOS JURÍDICOS DE LOS DATOS CATASTRALES. Retrieved from https://www.notariosyregistradores.com/doctrina/2011-catastro-efectos-juridicos.htm

Digital Guardian. (2018, April 06). What is GDPR (General Data Protection Regulation)? Understanding and Complying with GDPR Data Protection Requirements. Retrieved June 7, 2018, from https://digitalguardian.com/blog/what-gdpr-general-data-protection-regulation-understanding-and-complying-gdpr-data-protection

Freund, Y. (1996). Experiments with a New Boosting Algorithm. Retrieved May 8, 2018, from https://cseweb.ucsd.edu/~yfreund/papers/boostingexperiments.pdf

Fumo, D. (2017, June 15). Types of Machine Learning Algorithms You Should Know. Retrieved June 3, 2018, from https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861

Gobierno de España. (n.d.). Sede Electrónica del Catastro. Retrieved from http://www.sedecatastro.gob.es/

Gobierno de España. (n.d.). THE CADASTRAL ELECTRONIC SITE (SEC). Retrieved from http://www.catastro.meh.es/ayuda/english_ovc.htm

Idealista. (2017, October 17). Legal statement. Retrieved June 10, 2018, from https://www.idealista.com/en/info/nota-legal

Kaggle. (2017). Zillow Prize: Zillow's Home Value Prediction (Zestimate) | Kaggle. Retrieved from https://www.kaggle.com/c/zillow-prize-1/data

Klopp, W. (2016, June 05). How big data is impacting real estate. Retrieved June 13, 2018, from http://realestatethings.net/real-estate-big-data/

Koehrsen, W. (2018, March 03). Beyond Accuracy: Precision and Recall – Towards Data Science. Retrieved May 4, 2018, from https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c

Koehrsen, W. (2018, January 28). Overfitting vs. Underfitting: A Complete Example – Towards Data Science. Retrieved May 7, 2018, from https://towardsdatascience.com/overfitting-vs-underfitting-a-complete-example-d05dd7e19765

Ministerio de hacienda. (n.d.). Catastro. Retrieved from http://www.catastro.meh.es/esp/usos_utilidades.asp

Morgun, I. (2017, January 22). Classification using k-Nearest Neighbors in R. Retrieved May 6, 2018, from https://en.proft.me/2017/01/22/classification-using-k-nearest-neighbors-r/

Negm, A. (2015, December). A hypothetical example of Multilayer Perceptron Network. Retrieved May 7, 2018, from https://www.researchgate.net/figure/A-hypothetical-example-of-Multilayer-Perceptron-Network_fig4_303875065

Panthong, R. (2015, December 23). Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm. Retrieved May, 2018, from https://www.sciencedirect.com/science/article/pii/S1877050915035784

Patel, S. (2017, May 11). Chapter 3 : Decision Tree Classifier - Theory – Machine Learning 101 – Medium. Retrieved May 10, 2018, from https://medium.com/machine-learning-101/chapter-3-decision-trees-theory-e7398adac567

PricewaterhouseCoopers. (2017). Embracing the world of Data for Real Estate. Retrieved June 10, 2018, from https://www.pwc.lu/en/real-estate/docs/pwc-re-embracing-the-world-of-data.pdf

Rodríguez López, J. (2006, May). SITUACIÓN Y PERSPECTIVAS FUTURAS EN EL SECTOR INMOBILIARIO EN ESPAÑA. Retrieved May 2, 2018, from https://www.fomento.gob.es/NR/rdonlyres/CBEF58A7-1A35-43EF-97E8-D183D00B45E7/99227/SitPersp06.pdf

Rusin, M. (2018, February 2). What is meta-learning in machine learning? Retrieved May 8, 2018, from https://www.quora.com/What-is-meta-learning-in-machine-learning

Simonini, T. (2018, March 31). An introduction to Reinforcement Learning – freeCodeCamp. Retrieved June 6, 2018, from https://medium.freecodecamp.org/an-introduction-to-reinforcement-learning-4339519de419

Statista. (2018, February 28). Devices used to access the internet UK 2018 | Survey. Retrieved June 10, 2018, from https://www.statista.com/statistics/387447/consumer-electronic-devices-by-internet-access-in-the-uk/

StatSoft. (2013). Naive Bayes Classifier. Retrieved from http://www.statsoft.com/textbook/naive-bayes-classifier

Sunasra, M. (2017, November 11). Performance Metrics for Classification problems in Machine Learning. Retrieved May 6, 2018, from https://medium.com/greyatom/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b

Tasación. (n.d.). El precio de la vivienda en España - Tinsa - IMIE Interactivo. Retrieved from https://www.tinsa.es/precio-vivienda/

Vanschoren, J. (2013). OpenML. Retrieved May 2, 2018, from https://www.openml.org/a/estimation-procedures/1

Wu, E. (2015, February 3). What is the difference between a Bayesian network and an artificial neural network? Retrieved June 3, 2018, from https://www.quora.com/What-is-the-difference-between-a-Bayesian-network-and-an-artificial-neural-network

APPENDIX I: APPROVED CADASTRAL VALUES (2011)

Pol.	Denominación	MBC	Importe MBC	MBR	Importe MBR	VUB	Importe VUB	VRB	Importe VRB
046	SALAMANCA-CASTELLANA	1	700.00	1	1700.00			R08H	2905.00
045	SALAMANCA-LISTA	1	700.00	1	1700.00			R11J	2226.00
044	SALAMANCA-GUINDALERA	1	700.00	1	1700.00			R15Z	1587.00
043	SALAMANCA-FUENTE DEL BERRO	1	700.00	1	1700.00			R15L	1587.00
042	SALAMANCA-GOYA	1	700.00	1	1700.00			R10H	2440.00
041	SALAMANCA-RECOLETOS	1	700.00	1	1700.00			R08B	2905.00

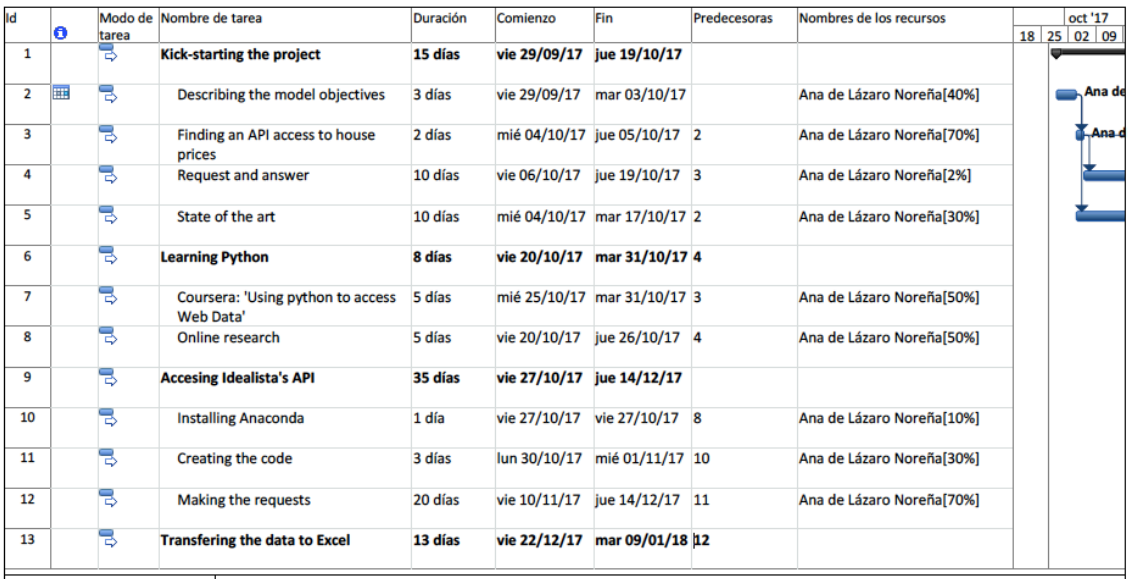
Polígono	Denominación	Nº Muestras	Uso	Valor catastral medio 2002 €/m²	Valor de Mercado medio €/m²	Valor catastral / Valor Mercado
041	Recoletos	4	R.Colectiva	985	5232	19,19
042	Goya	8	R.Colectiva	769	3628	21,16
043	Fuente del Berro	2	R.Colectiva	725	2873	25,25
044	Guindalera	17	R.Colectiva	771	2769	28,01
045	Lista	7	R.Colectiva	805	3578	22,67
046	Castellana	6	R.Colectiva	981	4730	20,77

TIPOLOGÍAS CONSTRUCTIVAS			CATEGORÍA								
USO	CLASE	MODALIDAD	1	2	3	4	5	6	7	8	9
1 RESIDENCIAL	1.1 VIVIENDAS COLECTIVAS de CARACTER URBANO	1.1.1 EDIFICACION ABIERTA	1,65	1,40	1,20	1,05	0,95	0,85	0,75	0,65	0,55
		1.1.2 EN MANZANA CERRADA	1,60	1,35	1,15	1,00	0,90	0,80	0,70	0,60	0,50
		1.1.3 GARAGES, TRASTEROS Y LOCALES EN ESTRUCTURA	0,80	0,70	0,62	0,53	0,46	0,40	0,30	0,26	0,20
	1.2 VIV. UNIFAMILIARES de CARACTER URBANO	1.2.1 EDIFICACION AISLADA O PAREADA	2,15	1,80	1,45	1,25	1,10	1,00	0,90	0,80	0,70
		1.2.2 EN LINEA O MANZANA CERRADA	2,00	1,65	1,35	1,15	1,05	0,95	0,85	0,75	0,65
		1.2.3 GARAGES Y PORCHES EN PLANTA BAJA	0,90	0,85	0,75	0,65	0,60	0,55	0,45	0,40	0,35
	1.3 EDIFICACION RURAL	1.3.1 USO EXCLUSIVO DE VIVIENDA	1,35	1,20	1,05	0,90	0,80	0,70	0,60	0,50	0,40
		1.3.2 ANEXOS	0,70	0,60	0,50	0,45	0,40	0,35	0,30	0,25	0,20
2 INDUSTRIAL	2.1 NAVES DE FABRICACION Y ALMACENAMIENTO	2.1.1 FABRICACION EN UNA PLANTA	1,05	0,90	0,75	0,60	0,50	0,45	0,40	0,37	0,35
		2.1.2 FABRICACION EN VARIAS PLANTAS	1,15	1,00	0,85	0,70	0,60	0,55	0,52	0,50	0,40
		2.1.3 ALMACENAMIENTO	0,85	0,70	0,60	0,50	0,45	0,35	0,30	0,25	0,20
	2.2 GARAGES Y APARCAMIENTOS	2.2.1 GARAJES	1,15	1,00	0,85	0,70	0,60	0,50	0,40	0,30	0,20
		2.2.2 APARCAMIENTOS	0,60	0,50	0,45	0,40	0,35	0,30	0,20	0,10	0,05
	2.3 SERVICIOS DE TRANSPORTE	2.3.1 ESTACIONES DE SERVICIO	1,80	1,60	1,40	1,25	1,20	1,10	1,00	0,90	0,80
		2.3.2 ESTACIONES	2,55	2,25	2,00	1,80	1,60	1,40	1,25	1,10	1,00
3 OFICINAS	3.1 EDIFICIO EXCLUSIVO	3.1.1 OFICINAS MULTIPLES	2,35	2,00	1,70	1,50	1,30	1,15	1,11	0,90	0,80
		3.1.2 OFICINAS UNITARIAS	2,55	2,20	1,85	1,60	1,40	1,25	1,10	1,00	0,90
	3.2 EDIFICIO MIXTO	3.2.1 UNIDO A VIVIENDAS	2,05	1,80	1,50	1,30	1,10	1,00	0,90	0,80	0,70
		3.2.2 UNIDO A INDUSTRIA	1,40	1,25	1,10	1,00	0,85	0,65	0,55	0,45	0,35
	3.3 BANCA Y SEGUROS	3.3.1 EN EDIFICIO EXCLUSIVO	2,95	2,65	2,35	2,10	1,90	1,70	1,50	1,35	1,20
		3.3.2 EN EDIFICIO MIXTO	2,65	2,35	2,10	1,90	1,70	1,50	1,35	1,20	1,05
4 COMERCIAL	4.1 COMERCIOS en EDIFICIO MIXTO	4.1.1 LOCALES COMERCIALES Y TALLERES	1,95	1,60	1,35	1,20	1,05	0,95	0,85	0,75	0,65
		4.1.2 GALERIAS COMERCIALES	1,85	1,65	1,45	1,30	1,15	1,00	0,90	0,80	0,70
	4.2 COMERCIOS en EDIFICIO EXCLUSIVO	4.2.1 EN UNA PLANTA	2,50	2,15	1,85	1,60	1,40	1,25	1,10	1,00	0,85
		4.2.2 EN VARIAS PLANTAS	2,75	2,35	2,00	1,75	1,50	1,35	1,20	1,05	0,90
	4.3 MERCADOS Y SUPERMERCADOS	4.3.1 MERCADOS	2,00	1,80	1,60	1,45	1,30	1,15	1,00	0,90	0,80
		4.3.2 HIPERMERCADOS Y SUPERMERCADOS	1,80	1,60	1,45	1,30	1,15	1,00	0,90	0,80	0,70
5 DEPORTES	5.1 CUBIERTOS	5.1.1 DEPORTES VARIOS	2,10	1,90	1,70	1,50	1,30	1,10	0,90	0,70	0,50
		5.1.2 PISCINAS	2,30	2,05	1,85	1,65	1,45	1,30	1,15	1,00	0,90
	5.2 DESCUBIERTOS	5.2.1 DEPORTES VARIOS	0,70	0,55	0,50	0,45	0,35	0,25	0,20	0,10	0,05
		5.2.2 PISCINAS	0,90	0,80	0,70	0,60	0,50	0,40	0,35	0,30	0,25
	5.3 AUXILIARES	5.3.1 VESTUARIOS, DEPURADORAS, CALEFACCION, etc..	1,50	1,35	1,20	1,05	0,90	0,80	0,70	0,60	0,50
	5.4 ESPECTACULOS DEPORTIVOS	5.4.1 ESTADIOS, PLAZAS DE TOROS	2,40	2,15	1,90	1,70	1,50	1,35	1,20	1,05	0,95
		5.4.2 HIPODROMOS, CANODROMOS, VELODROMOS, etc.	2,20	1,95	1,75	1,55	1,40	1,25	1,10	1,00	0,90
6 ESPECTACULOS	6.1 VARIOS	6.1.1 CUBIERTOS	1,90	1,70	1,50	1,35	1,20	1,05	0,95	0,85	0,75
		6.1.2 DESCUBIERTOS	0,80	0,70	0,60	0,55	0,50	0,45	0,40	0,35	0,30
	6.2 BARES MUSICALES SALAS DE FIESTAS DISCOTECAS	6.2.1 EN EDIFICIO EXCLUSIVO	2,65	2,35	2,10	1,90	1,70	1,50	1,35	1,20	1,05
		6.2.2 UNIDO A OTROS USOS	2,20	1,95	1,75	1,55	1,40	1,25	1,10	1,00	0,90
	6.3 CINES Y TEATROS	6.3.1 CINES	2,55	2,30	2,05	1,80	1,60	1,45	1,30	1,15	1,00
		6.3.2 TEATROS	2,70	2,40	2,15	1,90	1,70	1,50	1,35	1,20	1,05
7 OCIO Y HOSTELERIA	7.1 CON RESIDENCIA	7.1.1 HOTELES, HOSTALES, MOTELES	2,65	2,35	2,10	1,90	1,70	1,50	1,35	1,20	1,05
		7.1.2 APARTHOTELES, BUNGALOWS	2,85	2,55	2,30	2,05	1,85	1,65	1,45	1,30	1,15
	7.2 SIN RESIDENCIA	7.2.1 RESTAURANTES	2,60	2,35	2,00	1,75	1,50	1,35	1,20	1,05	0,95
		7.2.2 BARES Y CAFETERIAS	2,35	2,00	1,70	1,50	1,30	1,15	1,00	0,90	0,80
	7.3 EXPOSICIONES Y REUNIONES	7.3.1 CASINOS Y CLUBS SOCIALES	2,60	2,35	2,10	1,90	1,70	1,50	1,35	1,20	1,05
		7.3.2 EXPOSICIONES Y CONGRESOS	2,50	2,25	2,00	1,80	1,60	1,45	1,25	1,10	1,00
8 SANIDAD Y BENEFICENCIA	8.1 SANITARIOS CON CAMAS	8.1.1 SANATORIOS Y CLINICAS	3,15	2,80	2,50	2,25	2,00	1,80	1,60	1,45	1,30
		8.1.2 HOSPITALES	3,05	2,70	2,40	2,15	1,90	1,70	1,50	1,35	1,20
	8.2 SANITARIOS VARIOS	8.2.1 AMBULATORIOS Y CONSULTORIOS	2,40	2,15	1,90	1,70	1,50	1,35	1,20	1,05	0,95
		8.2.2 BALNEARIOS, CASAS DE BAÑOS	2,65	2,35	2,10	1,90	1,70	1,50	1,35	1,20	1,05
	8.3 BENEFICOS Y ASISTENCIA	8.3.1 CON RESIDENCIA (Asilos, Residencias, etc.)	2,45	2,20	2,00	1,80	1,60	1,40	1,25	1,10	1,00
		8.3.2 SIN RESIDENCIA (Comedores, Clubs, Guarderias, etc.)	1,95	1,75	1,55	1,40	1,25	1,10	1,00	0,90	0,80
9 CULTURALES Y	9.1 CULTURALES CON RESIDENCIA	9.1.1 INTERNADOS	2,40	2,15	1,90	1,70	1,50	1,35	1,20	1,05	0,95
		9.1.2 COLEGIOS MAYORES	2,60	2,35	2,10	1,90	1,70	1,50	1,35	1,20	1,05

Data source Appendix 1: Ministerio de hacienda. (2011). PONENCIA DE VALORES TOTAL DE BIENES INMUEBLES URBANOS DEL MUNICIPIO DE MADRID. Retrieved May 4, 2018, from http://www.catastro.minhap.es/ponencias/28/900/28900_PT_DOC1_2011.pdf

APPENDIX II: PROJECT PLANNING

Id	Modo de tarea	Nombre de tarea	Duración	Comienzo	Fin	Predecesoras	Nombres de los recursos
1		**Kick-starting the project**	**15 días**	**vie 29/09/17**	**jue 19/10/17**		
2		Describing the model objectives	3 días	vie 29/09/17	mar 03/10/17		Ana de Lázaro Noreña[40%]
3		Finding an API access to house prices	2 días	mié 04/10/17	jue 05/10/17	2	Ana de Lázaro Noreña[70%]
4		Request and answer	10 días	vie 06/10/17	jue 19/10/17	3	Ana de Lázaro Noreña[2%]
5		State of the art	10 días	mié 04/10/17	mar 17/10/17	2	Ana de Lázaro Noreña[30%]
6		**Learning Python**	**8 días**	**vie 20/10/17**	**mar 31/10/17**	**4**	
7		Coursera: 'Using python to access Web Data'	5 días	mié 25/10/17	mar 31/10/17	3	Ana de Lázaro Noreña[50%]
8		Online research	5 días	vie 20/10/17	jue 26/10/17	4	Ana de Lázaro Noreña[50%]
9		**Accesing Idealista's API**	**35 días**	**vie 27/10/17**	**jue 14/12/17**		
10		Installing Anaconda	1 día	vie 27/10/17	vie 27/10/17	8	Ana de Lázaro Noreña[10%]
11		Creating the code	3 días	lun 30/10/17	mié 01/11/17	10	Ana de Lázaro Noreña[30%]
12		Making the requests	20 días	vie 10/11/17	jue 14/12/17	11	Ana de Lázaro Noreña[70%]
13		**Transfering the data to Excel**	**13 días**	**vie 22/12/17**	**mar 09/01/18**	**12**	



Proyecto: Proyecto1
Fecha: mar 12/06/18

Tarea	Hito externo		Informe de resumen manual
División	Tarea inactiva		Resumen manual
Hito	Hito inactivo		Sólo el comienzo
Resumen	Resumen inactivo		Sólo fin
Resumen del proyecto	Tarea manual		Fecha límite
Tareas externas	Sólo duración		Progreso

Página 1

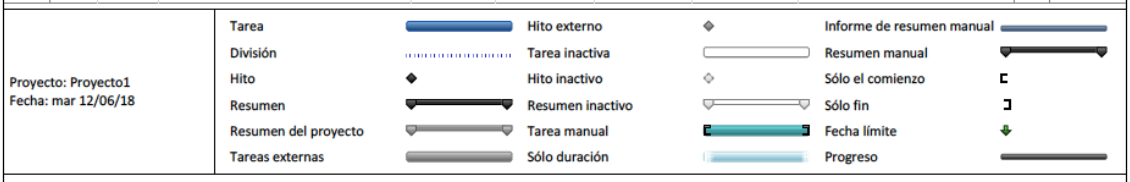
Id	Modo de tarea	Nombre de tarea	Duración	Comienzo	Fin	Predecesoras	Nombres de los recursos
14		From Spyder to Atom	4 días	vie 22/12/17	mié 27/12/17	11	Ana de Lázaro Noreña[50%]
15		From Atom to Excel	6 días	vie 22/12/17	mar 09/01/18	12	Ana de Lázaro Noreña[50%]
16		**Cadastral value**	**64 días**	**jue 02/11/17**	**mar 30/01/18**		
17		Research	6 días	jue 02/11/17	jue 09/11/17	11	Ana de Lázaro Noreña[70%]
18		Finding the values	5 días	vie 01/12/17	jue 07/12/17	17	Ana de Lázaro Noreña[70%]
19		Getting the data to Excel	4 días	vie 05/01/18	jue 11/01/18	18	Ana de Lázaro Noreña[60%]
20		Hypothesis and Calculations	2 días	lun 29/01/18	mar 30/01/18	19	Ana de Lázaro Noreña[70%]
21		**Security data, IPV**	**3 días**	**vie 12/01/18**	**mar 16/01/18**		
22		Finding the source	2 días	vie 12/01/18	lun 15/01/18	15	Ana de Lázaro Noreña[70%]
23		Getting the data to Excel	1 día	mar 16/01/18	mar 16/01/18	15;22	Ana de Lázaro Noreña[70%]
24		**Population, Parkland, Health, Restaurants, Social, Public Transport**	**14 días**	**vie 15/12/17**	**mié 03/01/18**		
25		Finding the source	5 días	vie 15/12/17	jue 21/12/17	12	Ana de Lázaro Noreña[70%]

Proyecto: Proyecto1
Fecha: mar 12/06/18

Tarea	Hito externo		Informe de resumen manual
División	Tarea inactiva		Resumen manual
Hito	Hito inactivo		Sólo el comienzo
Resumen	Resumen inactivo		Sólo fin
Resumen del proyecto	Tarea manual		Fecha límite
Tareas externas	Sólo duración		Progreso

Página 2

Id	Modo de tarea	Nombre de tarea	Duración	Comienzo	Fin	Predecesoras	Nombres de los recursos
26		Getting the data to Excel	5 días	jue 28/12/17	mié 03/01/18	12;25	Ana de Lázaro Noreña[70%]
27		**Cleaning the data**	**25 días**	**mié 31/01/18**	**mar 06/03/18**		
28		Cleaning the data	25 días	mié 31/01/18	mar 06/03/18	26;20;15	Ana de Lázaro Noreña[70%]
29		**Classification**	**51 días**	**mié 07/03/18**	**mié 16/05/18**		
30		Research	2 días	mié 07/03/18	jue 08/03/18	28	Ana de Lázaro Noreña[90%]
31		Distribution	8 días	vie 04/05/18	mié 16/05/18	30	Ana de Lázaro Noreña[70%]
32		**All data from Excel to Atom**	**3 días**	**vie 09/03/18**	**mar 13/03/18**		
33		New 'clean' Excel	1 día	vie 09/03/18	vie 09/03/18	28;30	Ana de Lázaro Noreña[70%]
34		Excel to Atom	2 días	lun 12/03/18	mar 13/03/18	33	Ana de Lázaro Noreña[70%]
35		**WAKA, classes**	**56 días**	**mié 14/03/18**	**mié 30/05/18**		
36		Formation on WEKA	12 días	mié 14/03/18	jue 29/03/18	34	Ana de Lázaro Noreña
37		Execution	5 días	vie 30/03/18	jue 05/04/18	36	Ana de Lázaro Noreña
38		Research on optimization	8 días	vie 06/04/18	mar 17/04/18	37	Ana de Lázaro Noreña



Id	Modo de tarea	Nombre de tarea	Duración	Comienzo	Fin	Predecesoras	Nombres de los recursos
39		Optimization	7 días	mar 22/05/18	mié 30/05/18	38	Ana de Lázaro Noreña
40		**WEKA, regression**	**18 días**	**mié 18/04/18**	**vie 11/05/18**		
41		Execution	8 días	mié 18/04/18	vie 27/04/18	36	Ana de Lázaro Noreña[80%]
42		Optimization	5 días	lun 30/04/18	vie 11/05/18	41	Ana de Lázaro Noreña[80%]
43		**Project documentation**	**42 días**	**mié 18/04/18**	**jue 14/06/18**		
44		Dissertation	30 días	mié 18/04/18	mar 12/06/18	26;34	Ana de Lázaro Noreña[20%]
45		Presentation	2 días	mié 13/06/18	jue 14/06/18	44	Ana de Lázaro Noreña[20%]